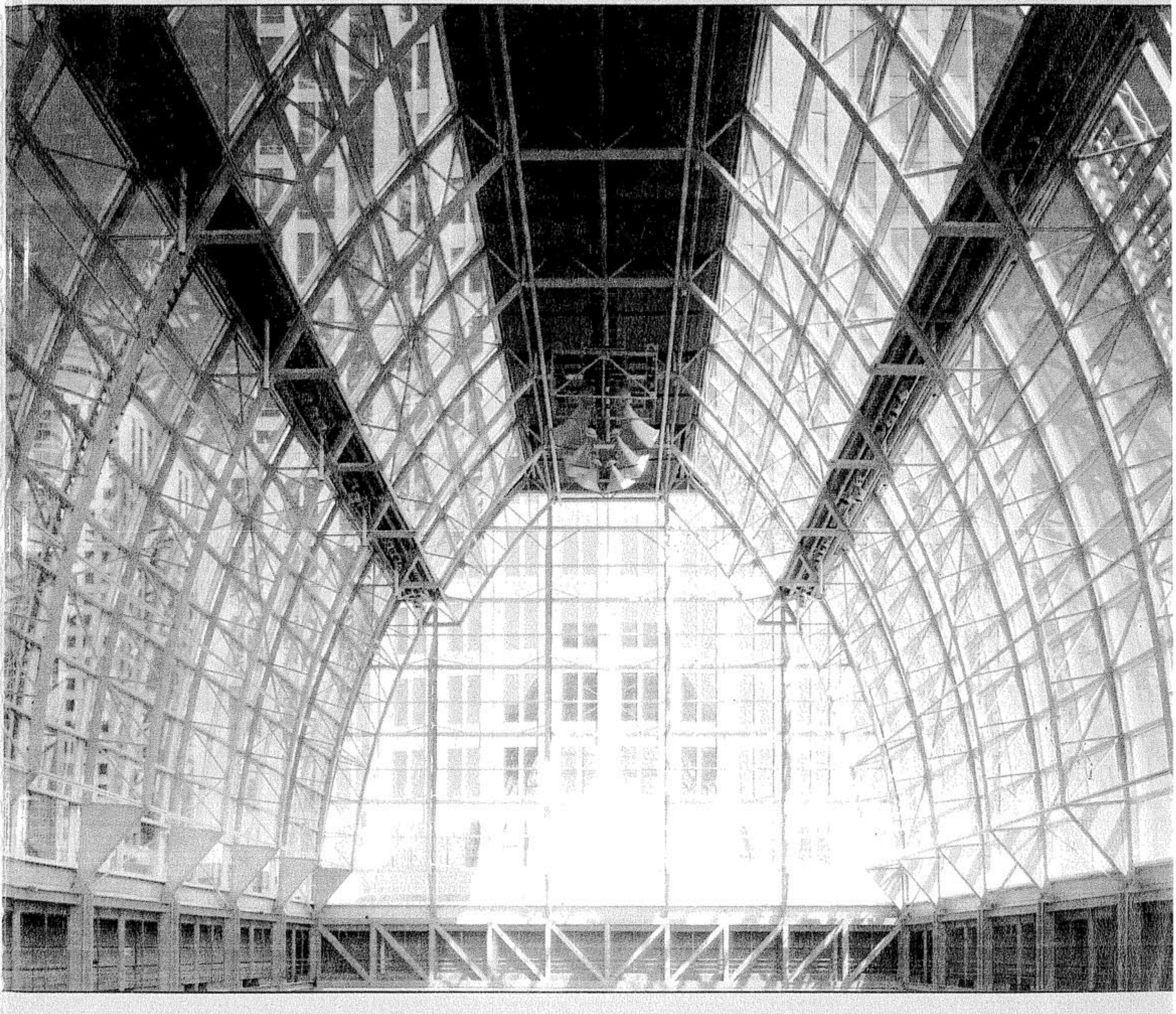
SECOND EDITION

ADVANCED ENGINEERING MATHEMATICS



Michael D. Greenberg

SELECTED FORMULAS

FIRST-ORDER LINEAR: $y' + p(x)y = q(x)$.

General solution $$y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right)$$

If $y(a) = b$, $$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right)$$

LEGENDRE EQUATION: $(1 - x^2)y'' - 2xy' + \lambda y = 0$

Bounded solutions $P_n(x)$ on $-1 \le x \le 1$ if $\lambda = n(n+1), \quad n = 0, 1, 2, \ldots$

BESSEL EQUATION: $x^2y'' + xy' + (x^2 - \nu^2)\,y = 0$

General solution $$y(x) = \begin{cases} AJ_\nu(x) + BY_\nu(x) \\ CH_\nu^{(1)}(x) + DH_\nu^{(2)}(x) \end{cases}$$

MODIFIED BESSEL EQUATION: $x^2y'' + xy' + (-x^2 - \nu^2)\,y = 0$

General solution $$y(x) = AI_\nu(x) + BK_\nu(x)$$

REDUCIBLE TO A BESSEL EQUATION: $\dfrac{d}{dx}\left(x^a \dfrac{dy}{dx}\right) + bx^c y = 0$

Solution $y(x) = x^{\nu/\alpha} Z_{|\nu|}\left(\alpha\sqrt{|b|}\,x^{1/\alpha}\right), \qquad \alpha = \dfrac{2}{c - a + 2}, \; \nu = \dfrac{1 - a}{c - a + 2},$
where $Z_{|\nu|}$ denotes $J_{|\nu|}$ and $Y_{|\nu|}$ if $b > 0$, and $I_{|\nu|}$ and $K_{|\nu|}$ if $b < 0$

MATRICES: $\mathbf{A}^{-1} = \dfrac{1}{\det\mathbf{A}}\,\mathrm{adj}\mathbf{A}, \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}, \quad (\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}}, \quad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$

CARTESIAN COORDINATES: $u = u(x, y, z), \quad \mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$

$$\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}, \quad d\mathbf{R} = dx\,\hat{\mathbf{i}} + dy\,\hat{\mathbf{j}} + dz\,\hat{\mathbf{k}},$$

$$dA = \begin{cases} dy\,dz & \text{(constant-}x\text{ surface)} \\ dx\,dz & \text{(constant-}y\text{ surface)} \\ dx\,dy & \text{(constant-}z\text{ surface)} \end{cases}, \qquad dV = dx\,dy\,dz$$

$$\nabla u = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right)u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

$$\nabla \cdot \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}$$

$$\nabla \times \mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_x & v_y & v_z \end{vmatrix}$$

CYLINDRICAL COORDINATES: $u = u(r, \theta, z), \;\; \mathbf{v} = v_r \hat{\mathbf{e}}_r + v_\theta \hat{\mathbf{e}}_\theta + v_z \hat{\mathbf{e}}_z$

$$x = r\cos\theta, \quad y = r\sin\theta, \quad z = z$$

$$\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z, \quad d\mathbf{R} = dr\hat{\mathbf{e}}_r + rd\theta\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z$$

$$dA = \begin{cases} r\,d\theta\,dz & \text{(constant-}r\text{ surface)} \\ dr\,dz & \text{(constant-}\theta\text{ surface)} \\ r\,dr\,d\theta & \text{(constant-}z\text{ surface)} \end{cases}, \qquad dV = r\,dr\,d\theta\,dz$$

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \quad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

$$\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta + \frac{\partial u}{\partial z}\hat{\mathbf{e}}_z$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}$$

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial}{\partial \theta}v_\theta + \frac{\partial}{\partial z}v_z$$

$$\nabla \times \mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\hat{\mathbf{e}}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\hat{\mathbf{e}}_\theta + \frac{1}{r}\left(\frac{\partial(rv_\theta)}{\partial r} - \frac{\partial v_r}{\partial \theta}\right)\hat{\mathbf{e}}_z$$

SPHERICAL COORDINATES: $u = u(\rho, \phi, \theta), \;\; \mathbf{v} = v_\rho \hat{\mathbf{e}}_\rho + v_\phi \hat{\mathbf{e}}_\phi + v_\theta \hat{\mathbf{e}}_\theta$

$$x = \rho\sin\phi\cos\theta, \quad y = \rho\sin\phi\sin\theta, \quad z = \rho\cos\phi$$

$$\mathbf{R} = \rho\hat{\mathbf{e}}_\rho, \quad d\mathbf{R} = d\rho\hat{\mathbf{e}}_\rho + \rho\,d\phi\hat{\mathbf{e}}_\phi + \rho\sin\phi\,d\theta\hat{\mathbf{e}}_\theta$$

$$dA = \begin{cases} \rho^2\,|\sin\phi|\,d\phi\,d\theta & \text{(constant-}\rho\text{ surface)} \\ \rho\,|\sin\phi|\,d\rho\,d\theta & \text{(constant-}\phi\text{ surface)} \\ \rho d\rho\,d\phi & \text{(constant-}\theta\text{ surface)} \end{cases}, \qquad dV = \rho^2\,|\sin\phi|\,d\rho\,d\phi\,d\theta$$

$$\begin{aligned} \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho} &= 0, & \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi} &= \hat{\mathbf{e}}_\phi, & \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta} &= \sin\phi\,\hat{\mathbf{e}}_\theta, \\ \frac{\partial\hat{\mathbf{e}}_\phi}{\partial\rho} &= 0, & \frac{\partial\hat{\mathbf{e}}_\phi}{\partial\phi} &= -\hat{\mathbf{e}}_\rho, & \frac{\partial\hat{\mathbf{e}}_\phi}{\partial\theta} &= \cos\phi\,\hat{\mathbf{e}}_\theta, \\ \frac{\partial\hat{\mathbf{e}}_\theta}{\partial\rho} &= 0, & \frac{\partial\hat{\mathbf{e}}_\theta}{\partial\phi} &= 0, & \frac{\partial\hat{\mathbf{e}}_\theta}{\partial\theta} &= -\sin\phi\,\hat{\mathbf{e}}_\rho - \cos\phi\,\hat{\mathbf{e}}_\phi \end{aligned}$$

$$\nabla u = \frac{\partial u}{\partial \rho}\,\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\frac{\partial u}{\partial \phi}\,\hat{\mathbf{e}}_\phi + \frac{1}{\rho\sin\phi}\frac{\partial u}{\partial \theta}\,\hat{\mathbf{e}}_\theta$$

$$\nabla^2 u = \frac{1}{\rho^2}\left[\frac{\partial}{\partial\rho}\left(\rho^2\frac{\partial u}{\partial\rho}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\,\frac{\partial u}{\partial\phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2}\right]$$

$$\nabla \cdot \mathbf{v} = \frac{1}{\rho^2}\frac{\partial}{\partial\rho}(\rho^2 v_\rho) + \frac{1}{\rho\sin\phi}\frac{\partial}{\partial\phi}(v_\phi\sin\phi) + \frac{1}{\rho\sin\phi}\frac{\partial v_\theta}{\partial\theta}$$

$$\nabla \times \mathbf{v} = \frac{1}{\rho\sin\phi}\left(\frac{\partial}{\partial\phi}(v_\theta\sin\phi) - \frac{\partial v_\phi}{\partial\theta}\right)\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\left(\frac{1}{\sin\phi}\frac{\partial v_\rho}{\partial\theta} - \frac{\partial(\rho v_\theta)}{\partial\rho}\right)\hat{\mathbf{e}}_\phi + \frac{1}{\rho}\left(\frac{\partial(\rho v_\phi)}{\partial\rho} - \frac{\partial v_\rho}{\partial\phi}\right)\hat{\mathbf{e}}_\theta$$

AREA ELEMENT: If $x = x(u,v),\ y = y(u,v),\ z = z(u,v):\quad dA = \sqrt{EG - F^2}\, du\, dv,$

$$E = x_u^2 + y_u^2 + z_u^2, \quad F = x_u x_v + y_u y_v + z_u z_v, \quad G = x_v^2 + y_v^2 + z_v^2$$

If $z = f(x,y):\quad dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy$

VOLUME ELEMENT: $dV = \left|\dfrac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\, dv\, dw = \left\|\begin{matrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{matrix}\right\| du\, dv\, dw$

LEIBNIZ RULE: $\dfrac{d}{dt}\displaystyle\int_{a(t)}^{b(t)} f(x,t)\, dx = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}\, dx + b'(t) f(b(t),t) - a'(t) f(a(t),t)$

FOURIER SERIES:

$f(x)$ 2ℓ-periodic: $\quad f(x) = a_0 + \displaystyle\sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right)$

$$a_0 = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, dx, \quad a_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \cos \frac{n\pi x}{\ell}\, dx, \quad b_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \sin \frac{n\pi x}{\ell}\, dx$$

$f(x)$ defined only on $0 < x < L$:

HRC: $\quad f(x) = a_0 + \displaystyle\sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L}, \qquad a_0 = \frac{1}{L} \int_0^L f(x)\, dx, \quad a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{L}\, dx$

HRS: $\quad f(x) = \displaystyle\sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L}, \qquad b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L}\, dx$

QRC: $\quad f(x) = \displaystyle\sum_{n=1,3,\ldots}^{\infty} a_n \cos \frac{n\pi x}{2L}, \qquad a_n = \frac{2}{L} \int_0^L f(x) \cos \frac{n\pi x}{2L}\, dx$

QRS: $\quad f(x) = \displaystyle\sum_{n=1,3,\ldots}^{\infty} b_n \sin \frac{n\pi x}{2L}, \qquad b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{2L}\, dx$

FOURIER INTEGRAL: $f(x) = \displaystyle\int_0^{\infty} [a(\omega) \cos \omega x + b(\omega) \sin \omega x]\, d\omega \qquad (-\infty < x < \infty)$

$$a(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(x) \cos \omega x, \quad b(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(x) \sin \omega x$$

FOURIER TRANSFORM: $F\{f(x)\} = \hat{f}(\omega) = \displaystyle\int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx$

$$F^{-1}\{\hat{f}(\omega)\} = f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}\, d\omega$$

LAPLACE TRANSFORM: $L\{f(t)\} = \overline{f}(s) = \displaystyle\int_0^{\infty} f(t) e^{-st}\, dt$

$$L^{-1}\{\overline{f}(s)\} = f(t) = \frac{1}{2\pi i} \int_{\gamma - i\infty}^{\gamma + i\infty} \overline{f}(s) e^{st}\, ds$$

DIVERGENCE THEOREM: $\int_{\mathcal{V}} \nabla \cdot \mathbf{v} \, dV = \int_S \hat{\mathbf{n}} \cdot \mathbf{v} \, dA$

GREEN'S FIRST IDENTITY: $\int_{\mathcal{V}} (\nabla u \cdot \nabla v + u \nabla^2 v) \, dV = \int_S u \frac{\partial v}{\partial n} \, dA$

GREEN'S SECOND IDENTITY: $\int_{\mathcal{V}} (u \nabla^2 v - v \nabla^2 u) \, dV = \int_S \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) dA$

STOKES'S THEOREM: $\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v} \, dA = \oint_C \mathbf{v} \cdot d\mathbf{R}$

GREEN'S THEOREM: $\int_S \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \oint_C P \, dx + Q \, dy$

STURM–LIOUVILLE EQUATION: $[p(x)y']' + q(x)y + \lambda w(x)y = 0$

SINE INTEGRAL FUNCTION: $Si(x) = \int_0^x \frac{\sin t}{t} \, dt, \qquad Si(\infty) = \frac{\pi}{2}$

EXPONENTIAL INTEGRAL FUNCTION: $E_1(x) = \int_x^\infty \frac{e^{-t}}{t} \, dt \quad (x > 0)$

GAMMA FUNCTION: $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, dt \quad (x > 0)$

ERROR FUNCTION: $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt, \quad \operatorname{erf}(\infty) = 1$

TRIGONOMETRIC FUNCTION IDENTITIES:

$$\cos x = \frac{e^{ix} + e^{-ix}}{2}, \quad \sin x = \frac{e^{ix} - e^{-ix}}{2i}$$
$$\cos(ix) = \cosh x, \quad \sin(ix) = i \sinh x$$
$$\cos^2 x + \sin^2 x = 1$$
$$\cos(A \pm B) = \cos A \cos B \mp \sin A \sin B$$
$$\sin(A \pm B) = \sin A \cos B \pm \sin B \cos A$$
$$\cos A \cos B = [\cos(A + B) + \cos(A - B)]/2$$
$$\sin A \cos B = [\sin(A + B) + \sin(A - B)]/2$$
$$\sin A \sin B = [\cos(A - B) - \cos(A + B)]/2$$

HYPERBOLIC FUNCTION IDENTITIES:

$$\cosh x = \frac{e^x + e^{-x}}{2}, \quad \sinh x = \frac{e^x - e^{-x}}{2}$$
$$\cosh(ix) = \cos x, \quad \sinh(ix) = i \sin x$$
$$\cosh^2 x - \sinh^2 x = 1$$
$$\cosh(A \pm B) = \cosh A \cosh B \pm \sinh A \sinh B$$
$$\sinh(A \pm B) = \sinh A \cosh B \pm \sinh B \cosh A$$

TAYLOR SERIES: $f(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!}(x - a)^2 + \frac{f'''(a)}{3!}(x - a)^3 + \cdots$

$$\frac{1}{1 - x} = 1 + x + x^2 + \cdots, \qquad |x| < 1 \quad \text{(Geometric Series)}$$
$$e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots, \qquad |x| < \infty$$
$$\cos x = 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 - \cdots, \qquad |x| < \infty$$
$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots, \qquad |x| < \infty$$

Advanced Engineering Mathematics

SECOND EDITION

Michael D. Greenberg
Department of Mechanical Engineering
University of Delaware, Newark, Delaware

PRENTICE HALL
Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Greenberg, Michael D., date–
Advanced engineering mathematics / Michael D. Greenberg. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-13-321431-1
1. Engineering mathematics. I. Title.
TA330.G725 1998
515'.14--dc21 97-43585
CIP

Technical Consultant: Dr. E. Murat Sozer

Acquisition editor: George Lobell
Editorial director: Tim Bozik
Editor-in-chief: Jerome Grant
Editorial assistant: Gale Epps
Executive managing editor: Kathleen Schiaparelli
Managing editor: Linda Mihatov Behrens
Production editor: Nick Romanelli
Director of creative services: Paula Maylahn
Art manager: Gus Vibal
Art director / cover designer: Jayne Conte
Cover photos: Timothy Hursley
Marketing manager: Melody Marcus
Marketing assistant: Jennifer Pan
Assistant vice president of production and manufacturing:
David Riccardi
Manufacturing buyer: Alan Fischer

Simon & Schuster / A Viacom Company
Upper Saddle River, New Jersey 07458

Printed in the United States of America.

ISBN 0-13-321431-1

Prentice-Hall International (UK) Limited, London
Prentice-Hall of Australia Pty. Limited, Sydney
Prentice-Hall Canada, Inc., Toronto
Prentice-Hall Hispanoamericana, S.A., Mexico
Prentice-Hall of India Private Limited, New Delhi
Prentice-Hall of Japan, Inc., Tokyo
Simon & Schuster Asia Pte, Ltd., Singapore
Editora Prentice-Hall do Brasil, Ltda., Rio de Janeiro

Curitiba
2001

Advanced
Engineering Mathematics

Contents

Part I: Ordinary Differential Equations

Part II: Linear Algebra

Part III: Scalar and Vector Field Theory

Part IV: Fourier Methods and Partial Differential Equations

Part V: Complex Variable Theory

Preface

Purpose and Prerequisites

This book is written primarily for a single- or multi-semester course in applied mathematics for students of engineering or science, but it is also designed for self-study and reference. By self-study we do not necessarily mean outside the context of a formal course. Even within a course setting, if the text can be read independently and understood, then more pedagogical options become available to the instructor.

The prerequisite is a full year sequence in calculus, but the book is written so as to be usable at both the undergraduate level and also for first-year graduate students of engineering and science. The flexibility that permits this broad range of use is described below in the section on Course Use.

Changes from the First Edition

Principal changes from the first edition are as follows:

1. **Part I on ordinary differential equations.** In the first edition we assumed that the reader had previously completed a first course in ordinary differential equations. However, differential equations is traditionally the first major topic in books on advanced engineering mathematics so we begin this edition with a seven chapter sequence on ordinary differential equations. Just as the book becomes increasingly sophisticated from beginning to end, these seven chapters are written that way as well, with the final chapter on nonlinear equations being the most challenging.
2. **Incorporation of a computer-algebra-system.** Several powerful computer environments are available, such as *Maple*, *Mathematica*, and MATLAB. We selected *Maple*, as a representative and user-friendly software. In addition to an Instructor's Manual, a brief student supplement is also available, which presents parallel discussions of *Mathematica* and MATLAB.
3. **Revision of existing material and format.** Pedagogical improvements that evolved through eight years of class use led to a complete rewriting rather than minor modifications of the text. The end-of-section exercises are refined and expanded.

Format

The book is comprised of five parts:

I Ordinary Differential Equations
II Linear Algebra
III Multivariable Calculus and Field Theory
IV Fourier Methods and Partial Differential Equations
V Complex Variable Theory

This breakdown is explicit only in the Contents, to suggest the major groupings of the chapters. Within the text there are no formal divisions between parts, only between chapters.

Each chapter begins with an introduction and (except for the first chapter) ends with a chapter review. Likewise, each section ends with a review called a closure, which is often followed by a section on computer software that discusses the *Maple* commands that are relevant to the material covered in that section; see, for example, pages 29–31. Subsections are used extensively to offer the instructor more options in terms of skipping or including material.

Course Use at Different Levels

To illustrate how the text might serve at different levels, we begin by outlining how we have been using it for courses at the University of Delaware: a sophomore/junior level mathematics course for mechanical engineers, and a first-year graduate level two-semester sequence in applied mathematics for students of mechanical, civil, and chemical engineering, and materials science. We denote these courses as U, G1, and G2, respectively.

Sophomore/junior level course (U). This course follows the calculus/differential equations sequence taught in the mathematics department. We cover three main topics:

Linear Algebra: Chapter 8, Sections 9.1–9.5 (plus a one lecture overview of Secs. 9.7–9.9), 10.1–10.6, and 11.1–11.3. The focus is n-space and applications, such as the mass-spring system in Sec. 10.6.2, Markov population dynamics in Sec. 11.2, and orthogonal modes of vibration in Sec. 11.3.

Field Theory: Chapters 14 and 16. The heart of this material is Chapter 16. Having skipped Chapter 15, we distribute a one page "handout" on the area element formula (18) in Sec. 15.5 since that formula is needed for the surface integrals that occur in Chapter 16. Emphasis is placed on the physical applications in the sections on the divergence theorem and irrotational fields since those applications lead to two of the three chief partial differential equations that will be studied in the third part of the course—the diffusion equation and the Laplace equation.

Fourier Series and PDE's: Sections 17.1–17.4, 18.1, 18.3, 18.6.1, 19.1–19.2.2, 20.1–20.3.1, 20.5.1–20.5.2. Solutions are by separation of variables, using only the half- and quarter-range Fourier series, and by finite differences.

First semester of graduate level course (G1). Text coverage is as follows: Sections 4.4–4.6, 5.1–5.6, Chapter 9, Secs. 11.1–11.4, 11.6, 13.5–13.8, 14.6, 15.4–15.6, Chapter 16, Secs. 17.3, 17.6–17.11, 18.1–18.3.1, 18.3.3–18.4, 19.1–19.2, 20.1–20.4. As in "U" we do cover the important Chapter 16, although quickly. Otherwise, the approach complements that in "U." For instance, in Chapter 9, "U" focuses on n-space, but "G1" focuses on generalized vector space (Sec. 9.6), to get ready for the Sturm–Liouville theory (Section 17.7); in Chapter 11 we emphasize the more advanced sections on diagonalization and quadratic forms, as well as Section 11.3.2 on the eigenvector expansion method in finite-dimensional space, so we can use that method to solve nonhomogeneous partial differential equations in later chapters. Likewise, in covering Chapter 17 we assume that the student has worked with Fourier series before so we move quickly, emphasizing the vector space approach (Sec. 17.6), the Sturm–Liouville theory, and the Fourier integral and transform. When we come to partial differential equations we use Sturm–Liouville eigenfunction expansions (rather than the half- and quarter-range formulas that suffice in "U"), integral transforms, delta functions, and Bessel and Legendre functions. In solving the diffusion equation in "U" we work only with the homogeneous equation and constant end conditions, but in "G1" we discuss the nonhomogeneous equation and nonconstant end conditions, uniqueness, and so on; these topics are discussed in the exercises.

Second semester of graduate level course (G2). In the second semester we complete the partial differential equation coverage with the methods of images and Green's functions, then turn to complex variable theory, the variational calculus, and an introduction to perturbation methods. For Green's functions we use a "handout," and for the variational calculus and perturbation methods we copy the relevant chapters from M.D. Greenberg, *Foundations of Applied Mathematics* (Englewood Cliffs, NJ: Prentice Hall, 1978). (If you are interested in using any of these materials please contact the College Mathematics Editor office at Prentice-Hall, Inc., One Lake Street, Upper Saddle River, NJ 07458.)

Text coverage is as follows: Chapters 21–24 on complex variable theory; then we return to PDE's, first covering Secs. 18.5–18.6, 19.3–19.4, and 20.3.2–20.4 that were skipped in "G1"; "handouts" on Green's functions, perturbation methods, and the variational calculus.

Shorter courses and optional Sections. A number of sections and subsections are listed as Optional in the Contents, as a guide to instructors in using this text for shorter or more introductory courses. In the chapters on field theory, for example, one could work only with Cartesian coordinates, and avoid the more difficult non-Cartesian case, by omitting those optional sections. We could have labeled the Sturm–Liouville theory section (17.7) as optional but chose not to, because it is such an important topic. Nonetheless, if one wishes to omit it, as we do in "U," that is possible, since subsequent use of the Sturm–Liouville theory in the PDE chapters is confined to optional sections and exercises.

Let us mention Chapter 4, in particular, since its development of series solutions, the method of Frobenius, and Legendre and Bessel functions might seem more detailed than you have time for in your course. One minimal route is to cover only Sections 4.2.2 on power series solutions of ordinary differential equations (ODE's) and 4.4.1 on Legendre polynomials, since the latter does not depend on the more detailed Frobenius material in Section 4.3. Then one can have Legendre functions available when the Laplace equation is studied in spherical coordinates. You might also want to cover Bessel functions but do not want to use class time to go through the Frobenius material. In my own course ("G1") I deal with Bessel functions by using a "handout" that is simpler and shorter, which complements the more thorough treatment in the text.

Exercises

Exercises are of different kinds and arranged, typically, as follows. First, and usually near the beginning of the exercise group, are exercises that follow up on the text or fill in gaps or relate to proofs of theorems stated in that section, thus engaging the student more fully in the reading (e.g., Exercises 1–3 in Section 7.2, Exercise 8 in Section 16.8). Second, there are usually numerous "drill type" exercises that ask the reader to mimic steps or calculations that are essentially like those demonstrated in the text (e.g., there are 19 matrices to invert by hand in Exercise 1 of Section 10.6, and again by computer software in Exercise 3).

Third, there are exercises that require the use of a computer, usually employing software that is explained at the end of the section or in an earlier section; these vary from drill type (e.g., Exercise 1, Section 10.6) to more substantial calculations (e.g., Exercise 15, Section 19.2). Fourth, there are exercises that involve physical applications (e.g., Exercises 8, 9, and 12 of Section 16.10, on the stream function, the entropy of an ideal gas, and integrating the equation of motion of fluid mechanics to obtain the Bernoulli equation). And, fifth, there are exercises intended to extend the text and increase its value as a reference book. In these, we usually guide the student through the steps so that the exercise becomes more usable for subsequent reference or self-study (e.g., see Exercises 17–22 of Section 18.3). Answers to selected exercises (which are denoted in the text by underlining the exercise number) are provided at the end of the book; a more complete set is available for instructors in the Instructor's Manual.

Specific Pedagogical Decisions

In Chapter 2 we consider the linear first-order equation and then the case of separable first-order equations. It is tempting to reverse the order, as some authors do, but we prefer to elevate the linear/nonlinear distinction, which grows increasingly important in engineering mathematics; to do that, it seems best to begin with the linear equation.

It is stated, at the beginning of Chapter 3 on linear differential equations of second order and higher, that the reader is expected to be familiar with the theory of the existence and uniqueness of solutions of linear algebraic equations, especially the role of the determinant of the coefficient matrix, even though this topic occurs later in the text. The instructor is advised to handle this need either by assigning, as prerequisite reading, the brief summary of the needed information given in Appendix B or, if a tighter blending of the differential equation and linear algebra material is desired, by covering Sections 8.1–10.6 before continuing with Chapter 3. Similarly, it is stated at the beginning of Chapter 3 that an elementary knowledge of the complex plane and complex numbers is anticipated. If the class does not meet that prerequisite, then Section 21.2 should be covered before Chapter 3. Alternatively, we could have made that material the first section of Chapter 3, but it seemed better to keep the major topics together—in this case, to keep the complex variable material together.

Some authors prefer to cover second-order equations in one chapter and then higher-order equations in another. My opinion about that choice is that: (i) it is difficult to grasp clearly the second-order case (especially insofar as the case of repeated roots is concerned) without seeing the extension to higher order, and (ii) the higher-order case can be covered readily, so that it becomes more efficient to cover both cases simultaneously.

Finally, let us explain why Chapter 8, on systems of linear algebraic equations and Gauss elimination, is so brief. Just as one discusses the real number axis before discussing functions that map one real axis to another, it seems best to discuss vectors before discussing matrices, which map one vector space into another. But to discuss vectors, span, linear dependence, bases, and expansions, one needs to know the essentials regarding the existence and uniqueness of solutions of systems of linear algebraic equations. Thus, Chapter 8 is intended merely to suffice until, having introduced matrices in Chapter 10, we can provide a more complete discussion.

Appendices

Appendix A reviews partial fraction expansions, needed in the application of Laplace and Fourier transforms. Appendix B summarizes the theory of the existence and uniqueness of solutions of linear algebraic equations, especially the role of the determinant of the coefficient matrix, and is a minimum prerequisite for Chapter 3. Appendices C through F are tables of transforms and conformal maps.

Instructor's Manual

An Instructor's Manual will be available to instructors from the office of the Mathematics Editor, College Department, Prentice-Hall, Inc., 1 Lake Street, Upper Saddle River, NJ 07458. Besides solutions to exercises, this manual contains additional pedagogical ideas for lecture material and some additional coverage, such as the Fast Fourier Transform, that can be used as "handouts."

Acknowledgements

I am grateful for a great deal of support in writing this second edition, but especially to Dr. E. Murat Sozer, who acted as a Technical Consultant. Dr. Sozer's work on the latex preparation of text and the figures went far beyond his original desire to learn the latex system and whose generous help was always extraordinary in quantity and quality. Sincere thanks also to my mathematics editor, George Lobell, for his insight and support, to Nick Romanelli in charge of production, to Anita Z. Hoover at the University of Delaware for extensive help with the latex computer system, and to these outside reviewers of the developing manuscript: Gregory Adams (Bucknell University), James M. W. Brownjohn (Nanyang Technical University), Melvin G. Davidson (Western Washington University), John H. Ellison (Grove City College), Mark Farris (Midwestern State University), Amitabha Ghosh (Rochester Institute of Technology), Evans M. Harrell, II (Georgia Tech.), Allen Hesse (Rochester Institute of Technology), Chung-yau Lam (Nanyang Technical University), Mohan Manoharan (Nanyang Technical University), James G. McDaniel (Boston University), Carruth McGehee (Lousiana State University), William Moss (Clemson University), Jean-Paul Nicol (Auburn University), John A. Pfaltzgraff (University of North Carolina, Chapel Hill), Mohammad Tavakoli (Chaffey College), David E. Weidner (University of Delaware), and Jingyi Zhu (University of Utah). I also thank these graduate students in this department for their help with working and debugging exercises: Rishikesh Bhalerao, Nadeem Faiz, Santosh Prabhu, and Rahul Rao and Zhaohui Chen.

I'm grateful to my wife, Yisraela, for her deep support and love when this task looked like more than I could handle, and for assuming many of my responsibilities, to give me the needed time. I dedicate this book to her.

Most of all, I am grateful to the Lord for bringing this book back to life and watching over all aspects of its writing and production: " From whence cometh my help? My help cometh from the Lord, who made heaven and earth." (Psalm 121)

Michael D. Greenberg

Chapter 1

Introduction to Differential Equations

1.1 Introduction

The mathematical formulation of problems in engineering and science usually leads to equations involving derivatives of one or more unknown functions. Such equations are called differential equations.

Consider, for instance, the motion of a body of mass m along a straight line, which we designate as an x axis. Let the mass be subjected to a force $F(t)$ along that axis, where t is the time. Then according to Newton's second law of motion

$$m\frac{d^2x}{dt^2} = F(t), \tag{1}$$

where $x(t)$ is the mass's displacement measured from the origin. If we prescribe the displacement $x(t)$ and wish to determine the force $F(t)$ required to produce that displacement, then the solution is simple: according to (1), we merely differentiate the given $x(t)$ twice and multiply by m.

However, if we know the applied force $F(t)$ and wish to determine the displacement $x(t)$ that results, then (1) is a "differential equation" on $x(t)$ since it involves the derivative, more precisely the second derivative, of the unknown function $x(t)$ with respect to t. To solve for x we need to "undo" the differentiations. That is, we need to integrate (1), twice in fact. For definiteness and simplicity, suppose that $F(t) = F_0$ is a constant. Then, integrating (1) once with respect to t gives

$$m\frac{dx}{dt} = F_0 t + A, \tag{2}$$

where A is an arbitrary constant of integration, and integrating again gives

$$mx = \frac{F_0}{2}t^2 + At + B,$$

or,

$$x(t) = \frac{1}{m}\left(\frac{F_0}{2}t^2 + At + B\right). \tag{3}$$

The constants of integration, A and B, can be found from (2) and (3) if the displacement x and velocity dx/dt are prescribed at the initial time ($t = 0$). If both $x(0)$ and $\frac{dx}{dt}(0)$ are zero, for instance, then (by setting $t = 0$) we find from (2) that $A = 0$, and then from (3) that $B = 0$. Thus, (3) gives the solution as $x(t) = F_0 t^2/2m$, and this solution holds for all $t \geq 0$.

Unfortunately, most differential equations cannot be solved this easily, that is, by merely undoing the derivatives. For instance, suppose that the mass is restrained by a coil spring that supplies a restoring force proportional to the displacement x, with constant of proportionality k (Fig. 1). Then in place of (1), the differential equation governing the motion is

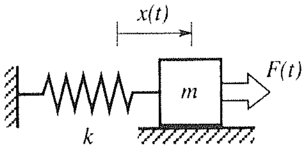


Figure 1. Mass/spring system.

$$m\frac{d^2x}{dt^2} = -kx + F(t)$$

or,

$$m\frac{d^2x}{dt^2} + kx = F(t). \tag{4}$$

After one integration, (4) becomes

$$m\frac{dx}{dt} + k\int x(t)\,dt = \int F(t)\,dt + A, \tag{5}$$

where A is the constant of integration. Since $F(t)$ is a prescribed function, the integral of $F(t)$ can be evaluated, but since $x(t)$ is the unknown, the integral of $x(t)$ cannot be evaluated, and we cannot proceed with our solution–by–integration.

Thus, we see that solving differential equations is not merely a matter of undoing the derivatives by direct integration. Indeed, the theory and technique involved is considerable, and will occupy us for these first seven chapters.

1.2 Definitions

In this section we introduce some of the basic terminology.

Differential equation. By a **differential equation** we mean an equation containing one or more derivatives of the function under consideration. Here are some examples of differential equations that we will study in later chapters:

$$m\frac{d^2x}{dt^2} + kx = F(t), \tag{1}$$

$$L\frac{d^2i}{dt^2} + \frac{1}{C}i = \frac{dE}{dt}, \tag{2}$$

$$\frac{d^2\theta}{dt^2} + \frac{g}{l}\sin\theta = 0, \tag{3}$$

$$\frac{dx}{dt} = cx, \tag{4}$$

$$\frac{d^2y}{dx^2} = C\sqrt{1 + \left(\frac{dy}{dx}\right)^2}, \tag{5}$$

$$EI\frac{d^4y}{dx^4} = -w(x). \tag{6}$$

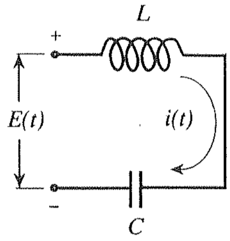


Figure 1. Electrical circuit, equation (2).

Equation (1) is the differential equation governing the linear displacement $x(t)$ of a body of mass m, subjected to an applied force $F(t)$ and a restraining spring of stiffness k, as mentioned in the preceding section.

Equation (2) governs the current $i(t)$ in an electrical circuit containing an inductor with inductance L, a capacitor with capacitance C, and an applied voltage source of strength $E(t)$ (Fig. 1), where t is the time.

Equation (3) governs the angular motion $\theta(t)$ of a pendulum of length l, under the action of gravity, where g is the acceleration of gravity and t is the time (Fig. 2).

Equation (4) governs the population $x(t)$ of a single species, where t is the time and c is a net birth/death rate constant.

Equation (5) governs the shape of a flexible cable or string, hanging under the action of gravity, where $y(x)$ is the deflection and C is a constant that depends upon the mass density of the cable and the tension at the midpoint $x = 0$ (Fig. 3).

Finally, equation (6) governs the deflection $y(x)$ of a beam subjected to a loading $w(x)$ (Fig. 4), where E and I are physical constants of the beam material and cross section, respectively.

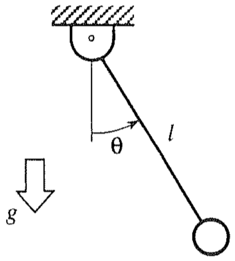


Figure 2. Pendulum, equation (3).

Ordinary and partial differential equations. We classify a differential equation as an **ordinary differential equation** if it contains ordinary derivatives with respect to a single independent variable, and as a **partial differential equation** if it contains partial derivatives with respect to two or more independent variables. Thus, equations (1)–(6) are ordinary differential equations (often abbreviated as **ODE**'s). The independent variable is t in (1)–(4) and x in (5) and (6).

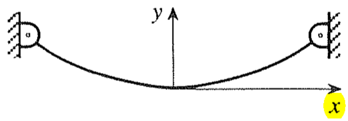


Figure 3. Hanging cable, equation (5).

Some representative and important partial differential equations (**PDE**'s) are as follows:

$$\alpha^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t}, \tag{7}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{8}$$

$$c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \frac{\partial^2 u}{\partial t^2}, \tag{9}$$

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = 0. \tag{10}$$

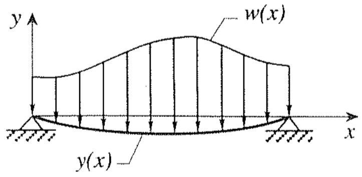


Figure 4. Loaded beam, equation (6).

Equation (7) is the *heat equation*, governing the time-varying temperature distribution $u(x, t)$ in a one-dimensional rod or slab; x locates the point under consideration within the material, t is the time, and α^2 is a material property called the diffusivity.

Equation (8) is the *Laplace equation*, governing the steady-state temperature distribution $u(x, y, z)$ within a three-dimensional body; x, y, z are the coordinates of the point within the material.

Equation (9) is the *wave equation*, governing the deflection $u(x, y, t)$ of a vibrating membrane such as a drum head.

Equation (10) is the *biharmonic equation*, governing the stream function $u(x, y)$ in the case of the slow (creeping) motion of a viscous fluid such as a film of wet paint.

Besides the possibility of having more than one independent variable, there could be more than one dependent variable. For instance,

$$\begin{aligned}\frac{dx_1}{dt} &= -(k_{21} + k_{31})x_1 + k_{12}x_2 + k_{13}x_3 \\ \frac{dx_2}{dt} &= k_{21}x_1 - (k_{12} + k_{32})x_2 + k_{32}x_3 \\ \frac{dx_3}{dt} &= k_{31}x_1 + k_{32}x_2 - (k_{13} + k_{23})x_3\end{aligned} \tag{11}$$

is a set, or system, of three ODE's governing the three unknowns $x_1(t), x_2(t), x_3(t)$; (11) arises in chemical kinetics, where x_1, x_2, x_3 are concentrations of three reacting chemical species, such as in a combustion chamber, where the k_{ij}'s are reaction rate constants, and where the reactions are, in chemical jargon, first-order reactions. Similarly,

$$\begin{aligned}\frac{\partial E_2}{\partial x} - \frac{\partial E_1}{\partial y} &= 0 \\ \frac{\partial E_1}{\partial x} + \frac{\partial E_2}{\partial y} &= \frac{\sigma(x, y)}{\epsilon}\end{aligned} \tag{12}$$

is a system of two PDE's governing the two unknowns $E_1(x, y)$ and $E_2(x, y)$, which are the x and y components of the electric field intensity, respectively, $\sigma(x, y)$ is the charge distribution density, and ϵ is the permittivity; these are the Maxwell's equations governing two-dimensional electrostatics.

At this point, we limit our subsequent attention to *ordinary* differential equations. We will not return to partial differential equations until much later on in this book. Thus, we will generally omit the adjective "ordinary," for brevity, and will speak only of "differential equations" over the next several chapters.

Order. We define the **order** of a differential equation as the order of the highest derivative therein. Thus, (4) is of first order, (1), (2), (3), and (5) are of second order, (6) is of fourth order, and (11) is a system of first-order ODE's.

More generally,

$$F\left(x, u(x), u'(x), u''(x), \ldots, u^{(n)}(x)\right) = 0 \tag{13}$$

is said to be an nth-order differential equation on the unknown $u(x)$, where n is the order of the highest derivative present in (13). Here, we use the standard prime notation for derivatives: $u'(x)$ denotes du/dx, $u''(x)$ denotes the second derivative, ... , and $u^{(n)}(x)$ denotes the nth derivative. In the fourth-order differential equation (6), for instance, in which the dependent variable is y rather than u, $F(x, y, y', y'', y''', y'''') = EIy'''' + w(x)$, which happens not to contain y, y', y'', or y'''.

Solution. A function is said to be a **solution** of a differential equation, over a particular domain of the independent variable, if its substitution into the equation reduces that equation to an identity everywhere within that domain.

EXAMPLE 1. The function $y(x) = 4 \sin x - x \cos x$ is a solution of the differential equation

$$y'' + y = 2 \sin x \tag{14}$$

on the entire x axis because its substitution into (14) yields

$$(-4 \sin x + 2 \sin x + x \cos x) + (4 \sin x - x \cos x) = 2 \sin x,$$

which is an identity for all x. Note that we said "a" solution rather than "the" solution since there are many solutions of (14):

$$y(x) = A \sin x + B \cos x - x \cos x \tag{15}$$

is a solution for *any* values of the constants A and B, as is verified by substitution of (15) into (14). [In a later chapter, we will be in a position to derive the solution (15), and to show that it is the most general possible solution, that is, that every solution of (14) can be expressed in the form (15).] ▮

EXAMPLE 2. The function $y(x) = 1/x$ is a solution of the differential equation

$$y' + y^2 = 0 \tag{16}$$

over any interval that does not contain the origin since its substitution into (16) gives $-1/x^2 + 1/x^2 = 0$, which relation is an identity, provided that $x \neq 0$. ▮

EXAMPLE 3. Whereas (14) admits an *infinity* of solutions [one for each choice of A and B in (15)], the equation

$$|y'| + |y| + 3 = 0 \tag{17}$$

evidently admits *none* since the two nonnegative terms and one positive term cannot possibly sum to zero for any choice of $y(x)$. ▮

In applications, however, one normally expects that if a problem is formulated carefully then it should indeed have a solution, and that the solution should be unique, that is, there should be one and only one. Thus, the issues of *existence* (Does the equation have any solution?) and *uniqueness* (If it does have a solution, is that solution unique?) are of important interest.

Initial-value problems and boundary-value problems. Besides the differential equation to be satisfied, the unknown function is often subjected to conditions at one or more points on the interval under consideration. Conditions specified at a single point (often the left end point of the interval), are called **initial conditions**, and the differential equation together with those initial conditions is called an **initial-value problem**. Conditions specified at both ends are called **boundary conditions**, and the differential equation together with the boundary conditions is called a **boundary-value problem**. For initial-value problems the independent variable is often the time, though not necessarily, and for boundary-value problems the independent variable is often a space variable.

EXAMPLE 4. *Straight-Line Motion of a Mass.* Consider once again the problem of predicting the straight-line motion of a body of mass m subjected to a force $F(t)$. According to Newton's second law of motion, the governing differential equation on the displacement $x(t)$ is $mx'' = F(t)$. Besides the differential equation, suppose that we wish to impose the conditions $x(0) = 0$ and $x'(0) = V$; that is, the initial displacement and velocity are 0 and V, respectively. Then the complete problem statement is the initial-value problem

$$\begin{aligned} mx''(t) &= F(t), \qquad (0 \le t < \infty) \\ x(0) &= 0, \quad x'(0) = V. \end{aligned} \tag{18}$$

That is, $x(t)$ is to satisfy the differential equation $mx'' = F(t)$ on the interval $0 \le t < \infty$ *and* the initial conditions $x(0) = 0$ and $x'(0) = V$. ▮

EXAMPLE 5. *Deflection of a Loaded Cantilever Beam.* Consider the deflection $y(x)$ of a cantilever beam of length L, under a loading $w(x)$ newtons per meter (Fig. 5). Using the so-called Euler beam theory, one finds that the governing problem is as follows:

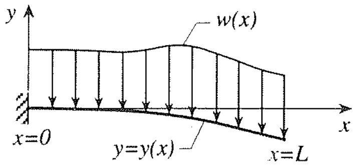


Figure 5. Loaded cantilever beam.

$$\begin{aligned} EIy'''' &= -w(x) \qquad (0 \le x \le L) \\ y(0) = 0, \quad y'(0) &= 0, \quad y''(L) = 0, \quad y'''(L) = 0, \end{aligned} \tag{19}$$

where E and I are known physical constants. The appended conditions are boundary conditions because some are specified at one end, and some at the other end, and (19) is therefore a boundary-value problem. The physical significance of the boundary conditions is as follows: $y(0) = 0$ is true simply by virtue of our chosen placement of the origin of the x, y coordinate system; $y'(0) = 0$ follows since the beam is cantilevered out of the wall, so that its slope at $x = 0$ is zero; $y''(L) = 0$ and $y'''(L) = 0$ because the "moment" and "shear force," respectively, are zero at the end of the beam. ▮

Linear and nonlinear differential equations. An nth-order differential equation is said to be **linear** if it is expressible in the form

$$\boxed{a_0(x)y^{(n)}(x) + a_1(x)y^{(n-1)}(x) + \cdots + a_n(x)y(x) = f(x),} \tag{20}$$

where $a_0(x), \ldots, a_n(x)$ are functions of the independent variable x alone, and **nonlinear** otherwise. Thus, equations (1), (2), (4), and (6) are linear, and (3) and (5) are nonlinear. If $f(x) = 0$, we say that (20) is **homogeneous**; if not, it is **nonhomogeneous**. If $a_0(x)$ does not vanish on the x interval of interest, then we may divide (20) by $a_0(x)$ (to normalize the leading coefficient) and re-express it as

$$y^{(n)}(x) + p_1(x)y^{(n-1)}(x) + \cdots + p_n(x)y(x) = q(x). \tag{21}$$

We will find that the theory of linear differential equations is quite comprehensive insofar as all of our major concerns – the existence and uniqueness of solutions, and how to *find* them, especially if the coefficients $a_0(x), \ldots, a_n(x)$ are constants. Even in the nonconstant coefficient case the theory provides substantial guidance.

Nonlinear equations are, in general, far more difficult, and the available theory is not nearly as comprehensive as for linear equations. Whereas for linear equations solutions can generally be found either in closed form or as infinite series, for nonlinear equations one might focus instead upon obtaining qualitative information about the solution, rather than the solution itself, or upon pursuing numerical solutions by computer simulation, or both.

The tendency in science and engineering, until around 1960, when high-speed digital computers became widely available, was to try to get by almost exclusively with linear theory. For instance, consider the nonlinear equation (3), namely,

$$\theta'' + \frac{g}{l}\sin\theta = 0, \tag{22}$$

governing the motion of a pendulum, where $\theta(t)$ is the angular displacement from the vertical and t is the time. If one is willing to limit one's attention to small motions, that is, where θ is small compared to unity (i.e., 1 radian), then one can use the approximation

$$\sin\theta = \theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \cdots \approx \theta$$

to replace the nonlinear equation (2) by the approximate "linearized" equation

$$\theta'' + \frac{g}{l}\theta = 0, \tag{23}$$

which (as we shall see in Chapter 3) is readily solved.

Unfortunately, the linearized version (23) is not only less and less accurate as larger motions are considered, it may even be incorrect in a qualitative sense as well. That is, from a phenomenological standpoint, replacing a nonlinear differential equation by an approximate linear one may amount to "throwing out the baby with the bathwater."

Thus, it is extremely important for us to keep the distinction between linear and nonlinear clearly in mind as we proceed with our study of differential equations. Aside from Sections 2.4 and 2.5, most of our study of nonlinear equations takes place in Chapters 6 and 7.

Closure. Notice that we have begun, in this section, to classify differential equations, that is, to categorize them by types. Thus far we have distinguished ODE's (ordinary differential equations) from PDE's (partial differential equations), established the order of a differential equation, distinguished initial-value problems from boundary-value problems, linear equations from nonlinear ones, and homogeneous equations from nonhomogeneous ones.

Why do we classify so extensively? Because the most general differential equation is far too difficult for us to deal with. The most reasonable program, then, is to break the set of all possible differential equations into various categories and to try to develop theory and solution strategies that are tailored to the specific nature of a given category. Historically, however, the early work on differential equations – by such mathematicians as *Leonhard Euler* (1707–1783), *Jakob* (James) *Bernoulli* (1654–1705) and his brother *Johann* (John) (1667–1748), *Joseph-Louis Lagrange* (1736–1813), *Alexis-Claude Clairaut* (1713–1765), and *Jean le Rond d'Alembert* (1717–1783) – generally involved attempts at solving specific equations rather than the development of a general theory.

From an applications point of view, we shall find that in many cases diverse physical phenomena are governed by the same differential equation. For example, consider equations (1) and (2) and observe that they are actually the same equation, to within a change in the names of the various quantities: $m \to L$, $k \to 1/C$, $F(t) \to dE(t)/dt$, and $x(t) \to i(t)$. Thus, to within these correspondences, their solutions are identical. We speak of the mechanical system and the electrical circuit as *analogs* of each other. This idea is deeper and more general than can be seen from this one example, and the result is that if one knows a lot about mechanical systems, for example, then one thereby knows a lot about electrical, biological, and social systems, for example, to whatever extent they are governed by differential equations of the same form.

Or, returning to PDE's for the moment, consider equation (7), which we introduced as the one-dimensional heat equation. Actually, (7) governs *any* one-dimensional diffusion process, be it the diffusion of heat by conduction, or the diffusion of material such as a pollutant in a river. Thus, when one is studying heat conduction one is also learning about all diffusion processes because the governing differential equation is the same. The significance of this fact can hardly be overstated as a justification for a careful study of the mathematical field of differential equations, or as a cause for marvel at the underlying design of the physical universe.

EXERCISES 1.2

1. Determine the order of each differential equation, and whether or not the given functions are solutions of that equation.

(a) $(y')^2 = 4y$; $y_1(x) = x^2$, $y_2(x) = 2x^2$, $y_3(x) = e^{-x}$

(b) $2yy' = 9\sin 2x$; $y_1(x) = \sin x$, $y_2(x) = 3\sin x$, $y_3(x) = e^x$

(c) $y'' - 9y = 0$; $y_1(x) = e^{3x} - e^x$, $y_2(x) = 3\sinh 3x$, $y_3(x) = 2e^{3x} - e^{-3x}$

(d) $(y')^2 - 4xy' + 4y = 0$; $y_1(x) = x^2 - x$, $y_2(x) = 2x - 1$

(e) $y'' + 9y = 0$; $y_1(x) = 4\sin 3x + 3\cos 3x$, $y_2(x) = 6\sin(3x + 2)$

(f) $y'' - y' - 2y = 6$; $y_1(x) = 5e^{2x} - 3$, $y_2(x) = -2e^{-x} - 3$

(g) $y''' - 6y'' + 12y' - 8y = 32 - 16x$; $y_1(x) = 2x - 1 + (A + Bx + Cx^2)e^{2x}$ for any constants A, B, C

(h) $y' + 2xy = 1$; $y_1(x) = Ae^{-x^2}\int_0^x e^{t^2}\,dt$, $y_2(x) = e^{-x^2}\int_a^x e^{t^2}\,dt$ for any constants A and a.

2. Verify that $u(x,t) = Ax + B + (C\sin\kappa x + D\cos\kappa x)\exp(-\kappa^2\alpha^2 t)$ is a solution of (7) for any constants A, B, C, D, κ. NOTE: We will sometimes use the notation exp( ) in place of $e^{(\ )}$ because it takes up less vertical space.

3. Verify that $u(x,y,z) = A\sin ax \sin by \sinh cz$ is a solution of (8) for any constants A, a, b, c, provided that $a^2 + b^2 = c^2$.

4. (a) Verify that $u(x,t) = (Ax + B)(Ct + D) + (E\sin\kappa x + F\cos\kappa x)(G\sin\kappa ct + H\cos\kappa ct)$ is a solution of the one-dimensional wave equation

$$c^2\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2},$$

for any constants $A, B, \dots, H, \kappa$.
(b) Verify that $u(x,t) = f(x - ct) + g(x + ct)$ is a solution of that equation for any twice-differentiable functions f and g.
(c) For what value(s) of the constant m is $u(x,t) = \sin(x + mt)$ a solution of that equation ?

5. For what value(s) of the constant λ will $y = \exp(\lambda x)$ be a solution of the given differential equation? If there are no such λ's, state that.

(a) $y' + 3y = 0$ (b) $y' + 3y^2 = 0$
(c) $y'' - 3y' + 2y = 0$ (d) $y'' - 2y' + y = 0$
(e) $y''' - y' = 0$ (f) $y''' - 2y'' - y' + 2y = 0$
(g) $y'''' - 6y'' + 5y = 0$ (h) $y'' + 5yy' + y = 0$

6. First, verify that the given function is a solution of the given differential equation, for any constants A, B. Then, solve for A, B so that y satisfies the given initial or boundary conditions.

(a) $y'' + 4y = 8x^2$; $y(x) = 2x^2 - 1 + A\sin 2x + B\cos 2x$; $y(0) = 1$, $y'(0) = 0$

(b) $y'' - y = x^2$; $y(x) = -x^2 - 2 + A\sinh x + B\cosh x$; $y(0) = -2$, $y'(0) = 0$

(c) $y'' + 2y' + y = 0$; $y(x) = (A + Bx)e^x$; $y(0) = 1$, $y(2) = 0$

(d) $y'' - y' = 0$; $y(x) = A + Be^x$; $y'(0) = 1$, $y(3) = 0$

7. Classify each equation as linear or nonlinear:

(a) $y' + e^x y = 4$ (b) $yy' = x + y$
(c) $e^x y' = x - 2y$ (d) $y' - \exp y = \sin x$
(e) $y'' + (\sin x)y = x^2$ (f) $y'' - y = \exp x$
(g) $yy''' + 4y = 3x$ (h) $y''' = y$

8. Recall that the nonlinear equation (5) governs the deflection $y(x)$ of the flexible cable shown in Fig. 3. Supposing that the sag is small compared to the span, suggest a linearized version of (5) that can be expected to give good accuracy in predicting the shape $y(x)$.

1.3 Introduction to Modeling

Emphasis in this book is on the mathematical analysis that begins once the problem has been formulated – that is, once the modeling phase has been completed. Detailed discussion of the modeling is handled best within applications courses, such as Heat Transfer, Fluid Mechanics, and Circuit Theory. However, we wish to emphasize the close relationship between the mathematics and the underlying physics, and to motivate the mathematics more fully. Thus, besides the purely mathematical examples in the text, we will include physical applications and some of the underlying modeling as well.

Our intention in this section is only to illustrate the nature of the modeling process, and we will do so through two examples. We suggest that you pay special attention to Example 1 because we will come back to it at numerous points later on in the text.

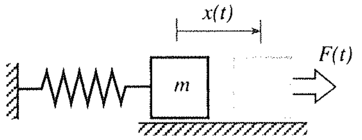


Figure 1. Mechanical oscillator.

EXAMPLE 1. *Mechanical Oscillator.* Consider a block of mass m lying on a table and restrained laterally by an ordinary coil spring (Fig. 1), and denote by x the displacement of the mass (measured as positive to the right) from its "equilibrium position;" that is, when $x = 0$ the spring is neither stretched nor compressed. We imagine the mass to be disturbed from its equilibrium position by an initial disturbance and/or an applied force $F(t)$, where t is the time, and we seek the differential equation governing the resulting displacement history $x(t)$.

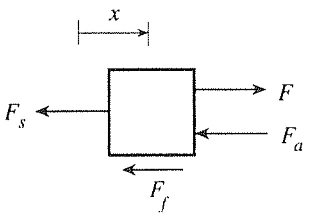


Figure 2. The forces, if $x > 0$ and $x' > 0$.

Our first step is to identify the relevant physics which, in this case, is Newton's second law of motion. Since the motion is constrained to be along a straight line, we need consider only the forces in the x direction, and these are shown in Fig. 2: F_s is the force exerted by the spring on the mass (the spring force, for brevity), F_a is the aerodynamic drag, F_f is the force exerted on the bottom of the mass due to its sliding friction, and F is the applied force, the driving force. How do we know if F_s, F_f, and F_a act to the left or to the right? The idea is to make assumptions on the signs of the displacement $x(t)$ and the velocity $x'(t)$ at the instant under consideration. For definiteness, suppose that $x > 0$ and $x' > 0$. Then it follows that each of the forces F_s, F_f, and F_a is directed to the left, as shown in Fig. 2. (The equation of motion that we obtain will be insensitive to those assumptions, as we shall see.) Newton's second law then gives

$$(\text{mass})(x \text{ acceleration}) = \text{sum of } x \text{ forces}$$

or,

$$mx'' = F - F_s - F_f - F_a, \tag{1}$$

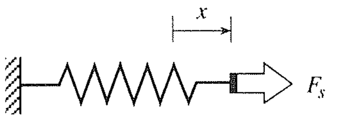


Figure 3. Spring force and displacement.

and we now need to express each of the forces F_s, F_f, and F_a in terms of the dependent and independent variables x and t.

Consider F_s first. If one knows enough about the geometry of the spring and the material of which it is made, one can derive an expression for F_s as a function of the extension x, as might be discussed in a course in Advanced Strength of Materials. In practice, however, one can proceed empirically, and more simply, by actually applying various positive (i.e., to the right, in the positive x direction) and negative forces (to the left) to the spring and measuring the resulting displacement x (Fig. 3). For a typical coil spring, the resulting graph will be somewhat as sketched in Fig. 4, where its steepening at A and B is due to the coils becoming completely compressed and completely extended, respectively. Thus, F_s in (1) is the function the graph of which is shown as the curve AB. (Ignore the dashed line L for the moment.)

Next, consider the friction force F_f. The modeling of F_f will depend upon the nature of the contact between the mass and the table – in particular, upon whether it is dry or lubricated. Let us suppose it is lubricated, so that the mass rides on a thin film of lubricant such as oil. To model F_f, then, we must consider the fluid mechanics of the lubricating film. The essential idea is that the stress τ (force per unit area) on the bottom of the mass is proportional to the gradient du/dy of the fluid velocity u (Fig. 5), where the constant of proportionality is the coefficient of viscosity μ: $\tau = \mu du/dy$. But $u(y)$ is found, in a course in Fluid Mechanics, to be a linear function, namely,

$$u(y) = \frac{u(h) - u(0)}{h} y = \frac{x'(t) - 0}{h} y = \frac{x'(t)}{h} y,$$

so

$$\tau = \mu \frac{du}{dy} = \frac{\mu}{h} x'(t).$$

Thus,

$$\begin{aligned} F_f &= (\text{stress } \tau)\ (\text{area } A \text{ of bottom of block}) \\ &= \left(\frac{\mu x'(t)}{h} \right) (A). \end{aligned}$$

That is, it is of the form

$$F_f = cx'(t), \tag{2}$$

for some constant c that we may consider as known. Thus, the upshot is that the friction force is proportional to the velocity. We will call c the *damping coefficient* because, as we will see in Chapter 3, the effect of the cx' term in the governing differential equation is to cause the motion to "damp out."

Likewise, one can model the aerodynamic drag force F_a, but let us neglect F_a on the tentative assumption that it can be shown to be small compared to the other two forces. Then (1) becomes

$$mx''(t) + cx'(t) + F_s(x) = F(t). \tag{3}$$

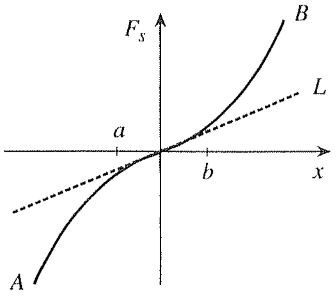


Figure 4. Force-displacement graph.

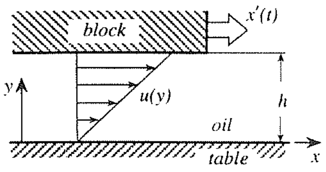


Figure 5. Lubricating film.

Equation (3) is nonlinear because $F_s(x)$ is not a linear function of x, as seen from its graph AB in Fig. 4. As a final simplifying approximation, let us suppose that the x motion is small enough, say between a and b in Fig. 4, so that we can linearize F_s, and hence the governing differential equation, by approximating the graph of F_s by its tangent line L. Since L is a straight line through the origin, it follows that we can express

$$F_s(x) \approx kx. \tag{4}$$

We call k the *spring stiffness*.

Thus, the final form of our governing differential equation, or equation of motion, is the linearized approximation

$$\boxed{mx'' + cx' + kx = F(t),} \tag{5}$$

on $0 \le t < \infty$, where the constants m, c, k and the applied force $F(t)$ are known. Equation (5) is important, and we will return to it repeatedly.

To this equation we wish to append suitable initial or boundary conditions. This particular problem is most naturally of initial-value type since we envision initiating the motion

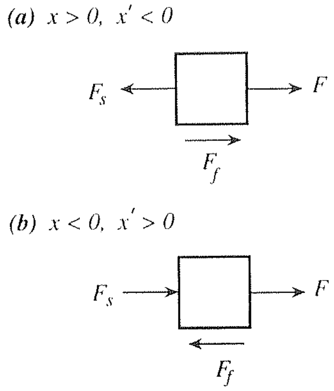


Figure 6. Other assumptions on the signs of x and x'.

in some manner at the initial time, say $t = 0$, and then considering the motion that results. Thus, to (5) we append initial conditions

$$x(0) = x_0 \quad \text{and} \quad x'(0) = x'_0, \tag{6}$$

for some (positive, negative, or zero) specified constants x_0 and x'_0. It should be plausible intuitively that we do need to specify both the initial displacement $x(0)$ and the initial velocity $x'(0)$ if we are to ensure a unique resulting motion. In any case, the theoretical appropriateness of the conditions (6) are covered in Chapter 3.

The differential equation (5) and initial conditions (6) comprise our resulting mathematical model of the physical system. By no means is there an exact correspondence between the model and the system since approximations were made in modeling the forces F_s and F_f, and in neglecting F_a entirely. Indeed, even our use of Newtonian mechanics, rather than relativistic mechanics, was an approximation.

This completes the modeling phase. The next step would be to solve the differential equation (5) subject to the initial conditions (6), for the motion $x(t)$.

COMMENT 1. Let us examine our claim that the resulting differential equation is insensitive to the assumptions made as to the signs of x and x'. In place of our assumption that $x > 0$ and $x' > 0$ at the instant under consideration, suppose we assume that $x > 0$ and $x' < 0$. Since $x > 0$, it follows that F_s acts to the left, and since $x' < 0$, it follows that F_f acts to the right. Then (Fig. 6a)

$$mx'' = F - F_s + F_f, \tag{7}$$

where we continue to neglect F_a. The sign of the F_f term is different in (7), compared with (1), because F_f now acts to the right, but notice that F_f now needs to be written as $F_f = c(-x'(t))$, rather than $cx'(t)$ since x' is negative. Further, F_s is still kx, so (7) becomes

$$mx'' = F(t) - kx + (-cx'), \tag{8}$$

which is indeed equivalent to (5), as claimed.

Next, what if $x < 0$ and $x' > 0$? This time (Fig. 6b)

$$mx'' = F + F_s - F_f, \tag{9}$$

which differs from (1) in the sign of the F_s term. But now F_s needs to be written as $F_s = k(-x(t))$ since x is negative. Further, F_f is cx', so (9) becomes

$$mx'' = F + k(-x) - cx',$$

which, again, is equivalent to (5). The remaining case, where $x < 0$ and $x' < 0$, is left for the exercises.

COMMENT 2. The approximation (4) was introduced from consideration of the graph shown in Fig. 4, but it amounts to expanding $F_s(x)$ in a Taylor series about the equilibrium point $x = 0$, as

$$F_s(x) = F_s(0) + F'_s(0)x + \frac{F''_s(0)}{2!}x^2 + \cdots$$

and linearizing – that is, cutting off after the first-order term:

$$\begin{aligned} F_s(x) &\approx F_s(0) + F'_s(0)x \\ &= 0 + kx = kx. \end{aligned}$$

This idea, the simplification of a differential equation by such tangent-line approximation, is of great importance in modeling.

COMMENT 3. The final equation for F_s, $F_s = kx$ is well known as **Hooke's law**, after *Robert Hooke* (1635–1703). Hooke published his law of elastic behavior in 1676 as the anagram *ceiiinosssttuv* and, two years later, the solution *ut tensio sic vis*: roughly, "as the force, so is the displacement." In view of the complexity with which we can now deal, Hooke's law must look quite modest, but one must appreciate it within its historical context. In spirit, it followed *Galileo Galilei* (1564-1642) who, in breaking lines established by the ancient Greeks, sought to establish a quantitative science, expressed in formulas and mathematical terms. For example, where Aristotle explained the increasing speed of a falling body in terms of the body moving more and more jubilantly as it approached its natural place (the center of the earth, which was believed to coincide with the center of the universe), Galileo sidestepped the question of cause entirely, and instead put forward the formula $v = 9.8t$, where v is the speed (in meters per second) and t is the time (in seconds). It may be argued that such departure from the less productive Greek tradition marked the beginning of modern science.

COMMENT 4. In science and engineering it is useful to think in terms of inputs and outputs. Here, there are three inputs, the two initial conditions and the applied force $F(t)$, and the output is the resulting solution, or response, $x(t)$. ∎

The foregoing introductory example illustrates several general truths about modeling. First, we see that it is not necessarily an easy task and generally requires a sound understanding of the underlying physics. Even in this little example one senses that obtaining suitable expressions for F_f and F_a (if one does include F_a) requires skillful handling of the fluid mechanics of the lubrication film and the aerodynamics of the moving block.

Second, we see that approximations will no doubt be necessary, and the idea is to make them judiciously. In this example we made several approximations. The expression $u(y) = x'(t)y/h$, for instance, is probably quite accurate but is not exact, especially near the ends of the block. Further, one can imagine that as the motion continues, the lubricant will heat up so that the viscosity μ will decrease. This effect is probably negligible, but we mention it in order to suggest that there is virtually no end to the levels of complexity that one may address, or suppress, in the modeling process. The key is to seek a level of complexity that will provide sufficient accuracy for the purpose at hand, and to seek a *uniform* level of approximation. For instance, it would hardly make sense to model F_f with great sophistication and accuracy if F_s is of comparable magnitude and is modeled only crudely.

To stay on this point a bit longer, note several more approximations that were implicit to our discussion. First, we implicitly assumed that the block is rigid, whereas the applied forces will cause some slight distortion of its shape and dimensions; again, neglect of this effect is surely an excellent approximation when considering the motion $x(t)$. Second, and more serious, notice that our empirical determination of $F_s(x)$ was based on a static test whereas, like the block, the spring is itself in motion. Thus, there is an inertial effect for the spring, analogous to the mx'' term for the mass, that we have neglected. If the mass of the spring is not negligible compared to that of the block, then that approximation may be insufficiently accurate.

Finally, notice carefully that we neglect a particular effect, in modeling, not on the grounds that it is small in an absolute sense, but because it is small relative to other effects. For instance, an aerodynamic force $F_a = 0.001$ newton may seem small numerically, but would not be negligible if F, F_s, and F_f were of comparable size.

Let us consider one more example.

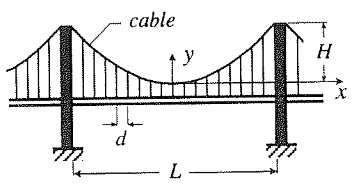


Figure 7. Suspension bridge cable.

EXAMPLE 2. *Suspension Bridge Cable.* To design a suspension bridge cable, one needs to know the relationships among the deflected shape of the cable, the tension in it, and the weight distribution that it needs to support.

In the case of a typical suspension bridge, the roadbed supported by the cables is much heavier than the cables themselves, so let us neglect the weight of the cables, and assume that the loading is due entirely to the roadbed. Consider the configuration shown schematically in Fig. 7.

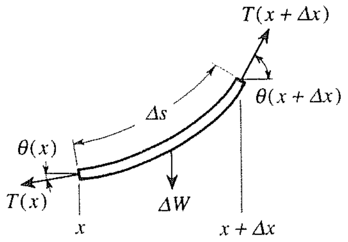


Figure 8. Typical cable element.

A cable is a distributed system, rather than one or more point masses, and for such systems a useful modeling approach is to isolate a typical element of the system and apply to it the relevant physical laws. In the present case a typical element is an incremental portion of the cable lying between x and $x + \Delta x$, for any x between $-L/2$ and $L/2$, as shown in Fig. 8: Δs is the arc length, T the tension in the cable, θ the inclination with respect to the horizontal, and ΔW the load supported by the element. If the roadbed is uniform and weighs $2w$ newtons per meter, then each of the two cables supports w newtons per meter, so $\Delta W = w\Delta x$.

Besides neglecting the weight of the cable itself, as mentioned above, there are three additional approximations that are implicit in the foregoing. First, in assuming a uniform load w per unit length, we have really assumed that the vertical support spacing d is very small compared to the span L, so that the intermittent support forces can be distributed as a uniform load. Second, in assuming that the tension is in the direction of the cable we have really assumed that the cable is flexible, a term that we now need to explain. The general state of affairs at any section, such as at the $x + \Delta x$ end of the element, is as shown in Fig. 9, namely, there can be a shear force V, a tensile force T through the centerline, and a moment or "couple" M. (V is the net effect of shearing stresses distributed over the face, and T and M are the net effect of normal stresses distributed over the face.) By a flexible string or cable, we mean one that is unable to support any shear V or moment M; that is, $V = M = 0$. For instance, if one takes a typical household string between the thumb and forefinger of each hand one finds that it offers virtually no resistance to shearing or bending, but considerable resistance to stretching. Thus, when we include only tensile forces in Fig. 8 we are assuming that the cable is flexible. Of course, if we imagine taking the suspension cables on the Golden Gate Bridge "between our fingers" we can imagine quite a bit of resistance to shearing and bending! But the point to keep in mind is that even those heavy cables would offer little resistance to shearing and bending by the enormous loads to which they are actually subjected by the weight of the roadbed.

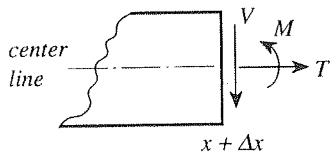


Figure 9. Forces and moments at an end.

Finally, we assume that the cable is inextensible, even under the large tensions that are anticipated.

If we can accept these assumptions we can now apply the relevant physics which, again, is Newton's second law of motion. But this time there is no acceleration, so the element is in static equilibrium. Thus, the x and y forces must each sum to zero:

$$x: \quad T(x+\Delta x)\cos\theta(x+\Delta x) - T(x)\cos\theta(x) = 0, \tag{10a}$$
$$y: \quad T(x+\Delta x)\sin\theta(x+\Delta x) - T(x)\sin\theta(x) - w\Delta x = 0. \tag{10b}$$

Dividing each of these equations by Δx and letting $\Delta x \to 0$, we obtain

$$\frac{d}{dx}(T\cos\theta) = 0, \tag{11a}$$

$$\frac{d}{dx}(T\sin\theta) = w, \tag{11b}$$

or, upon integration,

$$T\cos\theta = A, \tag{12a}$$

$$T\sin\theta = wx + B, \tag{12b}$$

where A, B are arbitrary constants. Dividing (12b) by (12a), to eliminate the unknown tension $T(x)$, and noting that $\tan\theta = dy/dx$, we obtain the differential equation

$$\frac{dy}{dx} = \frac{w}{A}x + \frac{B}{A} \tag{13}$$

governing the cable shape $y(x)$, which we are trying to determine.

In this case, the solution is simple enough so that we might as well complete the solution. To solve, we merely integrate (13), obtaining

$$y(x) = \frac{w}{2A}x^2 + \frac{B}{A}x + C,$$

where A, B, C are arbitrary constants. To evaluate them, we invoke the associated boundary conditions:

$$y(0) = 0 \quad \text{(by choice of location of origin)}, \tag{14a}$$
$$y'(0) = 0 \quad \text{(by symmetry about x = 0)}, \tag{14b}$$
$$y\left(\frac{L}{2}\right) = H \quad \text{(from Fig. 7)}. \tag{14c}$$

Equation (14a) gives $C = 0$, and (14b) gives $B/A = 0$, and hence $B = 0$. Thus far,

$$y(x) = \frac{w}{2A}x^2,$$

and (14c) then gives $A = wL^2/8H$. Thus, the cable's shape is given by the parabola

$$y(x) = \frac{4H}{L^2}x^2. \tag{15}$$

Finally, the distribution of tension $T(x)$ may be found by squaring and adding (12a) and (12b):

$$\begin{aligned} T(x) &= \sqrt{A^2 + (wx+B)^2} = \sqrt{\left(\frac{wL^2}{8H}\right)^2 + w^2x^2} \\ &= w\sqrt{x^2 + \frac{L^4}{64H^2}}. \end{aligned} \tag{16}$$

In a sense, obtaining $y(x)$ and $T(x)$ marks the end of the analysis, and if we are content that the expressions (15) and (16) are sufficiently accurate, then the next step would be to use them to help in the actual bridge design. Before doing that, however, we should check those results, for there may be errors in our analysis. Also, the approximations that we have made may prove to be insufficiently accurate.

One of the standard ways of checking results is by means of special cases and limiting cases, for which the correct results are known or obvious. Considering (15), we observe first that the parabolic shape looks *reasonable*. Furthermore, (15) implies that $y(x) \to 0$, over $-L/2 \leq x \leq L/2$, as $H \to 0$ with L fixed, and also that $y(x) \to 0$ at each x, as $L \to \infty$ with H fixed. These results look reasonable too. Turning to (16), observe that the tension becomes infinite throughout the cable as $H \to 0$, as expected. (Try straightening out a loaded washline by pulling on one end!) Finally, consider the limiting case $H \to \infty$, with L fixed. In that case, (16) gives $T(L/2) \to wL/2$, which agrees with the result obtained from a simple consideration of the physics (Exercise 2). ▮

Closure. The purpose of this section is to illustrate the modeling process, whereby one begins with the physical problem at hand and ends up with an equivalent mathematical statement, or model. Actually, we should say approximately equivalent since the modeling process normally involves a number of assumptions and approximations. By no means do we claim to show how to model in general, but only to illustrate the modeling process and the intimate connection between the physics and the mathematics. As we proceed through this text we will attempt to keep that connection in view, even though our emphasis will be on the mathematics.

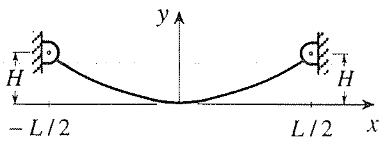
Finally, let us note that when we speak of the physical problem and the physics we intend those words to cover a much broader range of possibilities. For instance, the problem might be in the realm of economics, such as predicting the behavior of the Dow Jones Stock Index as a function of time. In that case the relevant "physical laws" would be economic laws such as the law of supply and demand. Or, the problem might fall in the realm of sociology, ecology, biology, chemistry, and so on. In any case, the general modeling approach is essentially the same, independent of the field of application.

EXERCISES 1.3

1. In Example 1 we showed that the same differential equation, (5), results, independent of whether $x > 0$ and $x' > 0$, or $x > 0$ and $x' < 0$ or $x < 0$ and $x' > 0$. Consider the last remaining case, $x < 0$ and $x' < 0$, and show that, once again, one obtains equation (5).

2. At the end of Example 2, we stated that the result $T(L/2) \to wL/2$, obtained from (16), for the limiting case where $H \to \infty$ with L fixed, agrees with the result obtained from a simple consideration of the physics. Explain that statement.

3. (*Catenary*) In our Suspension Bridge Cable example we neglected the weight of the cable itself relative to the weight of the roadbed. At the other extreme, suppose that the weight of the roadbed (or other loading) is negligible compared to the weight of the cable. Indeed, consider a uniform flexible cable, or catenary, hanging under the action of its own weight only, as sketched in the figure. Then Fig. 8 still holds, but with $\Delta W = \mu \Delta s$, where μ is the weight per unit arc length of the cable.



(a) Proceeding somewhat as in (10)–(12), derive the governing differential equation

$$y'' = C\sqrt{1 + y'^2}, \tag{3.1}$$

where C is an unknown constant.

(b) Since $y(x)$ is symmetric about $x = 0$, it suffices to consider the interval $0 \le x \le L/2$. Then we have the boundary conditions $y(0) = 0$, $y'(0) = 0$, and $y(L/2) = H$. Verify (you need not derive it) that

$$y(x) = \frac{1}{C}(\cosh Cx - 1) \tag{3.2}$$

satisfies (3.1) and the boundary conditions $y(0) = 0$ and $y'(0) = 0$. But it remains to determine C. Invoking the remaining boundary condition, $y(L/2) = H$, show that C satisfies the equation

$$H = \frac{1}{C}\left(\cosh \frac{CL}{2} - 1\right). \tag{3.3}$$

Unfortunately, (3.3) is a transcendental equation for C, so that we cannot solve it explicitly. We can solve it numerically, for given values of H and L, but you need not do that.

(c) As a partial check on these results, notice that they should reduce to the parabolic cable solution in the limiting case where the sag-to-span ratio H/L tends to zero, for then the load per unit x length, due to the weight of the cable, approaches a constant, as it is in Example 2, where the load is due entirely to the uniform roadbed. The problem that we pose for you is to carry out that check. HINT: Think of L as fixed and H tending to zero. For H to approach zero, in (3.3), we need $CL/2$ to approach zero – that is, $C \to 0$. Thus, we can expand the $\cosh Cx - 1$ in (3.2) in a Maclaurin series in C and retain the leading term. Show that that step gives $y(x) \approx Cx^2/2$, and the boundary condition $y(L/2) = H$ enables us to determine C. The result should be identical to (15).

(d) Actually, for small sag-to-span ratio we should be able to neglect the y'^2 term in (3.1), relative to unity, so that (3.1) can be linearized as

$$y'' = C. \tag{3.4}$$

Integrating (3.4) and using the boundary conditions $y(0) = 0$, $y'(0) = 0$, and $y(L/2) = H$, show that one obtains (15) once again.

Chapter 2

Differential Equations of First Order

2.1 Introduction

In studying algebraic equations, one considers the very simple first-degree polynomial equation $ax = b$ first, then the second-degree polynomial equation (quadratic equation), and so on. Likewise, in the theory of differential equations it is reasonable and traditional to begin with first-order equations, and that is the subject of this chapter. In Chapter 3 we turn to equations of second order and higher.

Recall that the general first-order equation is given by

$$F(x, y, y') = 0, \tag{1}$$

where x and y are independent and dependent variables, respectively. In spite of our analogy with algebraic equations, first-order differential equations can fall anywhere in the spectrum of complexity, from extremely simple to hopelessly difficult. Thus, we identify several different subclasses of (1), each of which is susceptible to a particular solution method, and develop them in turn. Specifically, we consider these subclasses: the linear equation $a_0(x)y' + a_1(x)y = f(x)$ in Section 2.2, "separable" equations in Section 2.4, equations that are "exact" (or can be made exact) in Section 2.5, and various other more specialized cases within the exercises.

These subclasses are not mutually exclusive. For instance, a given equation could be both linear and separable, in which case we could solve it by one method or the other. Given such a choice, choose whichever you prefer. In other cases the given equation might not fit into *any* of these subclasses and might be hopelessly difficult from an analytical point of view. Thus, it will be important to complement our analytical methods with numerical solution techniques. But that is a long way off, in Chapter 6. Analytical methods and the general theory of differential equations will occupy us in Chapters 2 through 5.

It should be stressed that the equation types that are susceptible to the analytical solution techniques described in these chapters can also be solved analytically by computer algebra software that is currently available, such as *Maple*, *Mathematica*, and *MATLAB*, and this approach is discussed here as well. One needs to know both: the underlying theory and solution methodology on one hand, and the available computer software on the other.

2.2 The Linear Equation

The first case that we consider is the general first-order linear differential equation

$$a_0(x)y' + a_1(x)y = f(x). \tag{1}$$

Dividing through by $a_0(x)$ [which is permissible if $a_0(x) \neq 0$ over the x interval of interest], we can re-express (1) in the more concise form

$$\boxed{y' + p(x)y = q(x).} \tag{2}$$

We assume that $p(x)$ and $q(x)$ are continuous over the x interval of interest.

2.2.1. Homogeneous case. It is tempting to think that to solve (2) for $y(x)$ we need to get rid of the derivative, and that we can accomplish that merely by integration. It's true that if we integrate (2) with respect to x,

$$\int y'\,dx + \int p\,y\,dx = \int q\,dx,$$

then the first term reduces nicely to y (plus a constant of integration), but the catch is that the $\int p\,y\,dx$ term becomes a stumbling block because we cannot evaluate it since $y(x)$ is unknown! Essentially, then, we have merely converted the differential equation (2) to an "integral equation" – that is, one involving the integral of the unknown function. Thus, we are no better off.

To solve (2), we begin with the simpler special case where $q(x)$ is zero,

$$y' + p(x)y = 0, \tag{3}$$

which is called the **homogeneous** version of (2). To solve (3), divide by y (assuming that y is nonzero on the x interval of interest. This assumption is tentative since y is not yet known) and integrate on x. Using the fact that $y'dx = dy$, from the calculus, we thus obtain

$$\int \frac{dy}{y} + \int p(x)\,dx = 0, \tag{4}$$

and recalling that

$$\int \frac{dx}{x} = \ln|x| + \text{constant}, \tag{5}$$

(4) gives

$$\ln|y| = -\int p(x)\,dx + C,$$

where the arbitrary constant C can include the arbitrary constant of integration from the p integral as well. Thus,

$$|y(x)| = e^{-\int p(x)\,dx + C} = e^C e^{-\int p(x)\,dx} = Be^{-\int p(x)\,dx}, \tag{6}$$

where we have set $e^C \equiv B$ for simplicity. Since C is real, e^C is nonnegative, so B is an arbitrary nonnegative constant: $B \geq 0$. The integral $\int p(x)\,dx$ does indeed exist since we have assumed that $p(x)$ is continuous. Finally, it follows from (6) that $y(x) = \pm B \exp(-\int p\,dx)$ or,

$$\boxed{y(x) = Ae^{-\int p(x)\,dx}} \tag{7}$$

if we now allow the arbitrary constant A to be positive, zero, or negative.

Observe that our tentative assumption, just below (3), that $y(x) \neq 0$, is now seen to be justified because the exponential function in (7) is strictly positive. [Of course $y(x) = 0$ if $A = 0$, but in that simple case $y(x) = 0$ is seen to satisfy (3) without further ado.] Summarizing, the solution of the homogeneous equation (3) is given by (7), where A is an arbitrary constant.

The presence of the arbitrary constant A, in (7), permits the solution (7) to satisfy an initial condition, if one is specified. Thus, suppose that we seek a solution $y(x)$ of our differential equation $y' + p(x)y = 0$ that satisfies an initial condition $y(a) = b$, for specified values of a and b. For this purpose, it is convenient to re-express (7) as

$$y(x) = Ae^{-\int_a^x p(\xi)\,d\xi}, \tag{8}$$

which is equivalent to (7) since $\int p(x)\,dx$ and $\int_a^x p(\xi)\,d\xi$ differ at most by an additive constant, say D, and the resulting e^D factor in (8) can be absorbed into the arbitrary constant A.

A point of notation: why do we change the integration variable from x to ξ in (8)? Because $\int_a^x p(\xi)\,d\xi$ means that we integrate along some axis, say a ξ axis, from a to x. Thus, x is a fixed endpoint, and is different from the integration variable that runs from a to x. To write $\int_a^x p(x)\,dx$ runs the risk of confusing the x's inside the integral with the fixed endpoint x. The *name* of the integration variable is immaterial, so one calls it a "dummy variable." For instance, $\int_0^x \xi\,d\xi$, $\int_0^x \eta\,d\eta$, and $\int_0^x \rho\,d\rho$ are all the same, namely, $x^2/2$. One often sees the letter ξ used as a dummy x variable because it is the Greek version of x. In Roman letters it is written as xi and is pronounced as ksē. Occasionally, we may be guilty of bad notation and write an integral as $\int_a^x f(x)\,dx$, simply to minimize the letters used, but even then we need to remember the distinction between the x in the upper limit and the x's inside the integral. In fact, this notation is typical in books on engineering and science, where there is less focus on such points of rigor.

Now imposing the condition $y(a) = b$ on (8) gives $y(a) = b = Ae^0 = A$, so $A = b$ and hence

$$\boxed{y(x) = be^{-\int_a^x p(\xi)\,d\xi}.} \tag{9}$$

Thus, (9) satisfies the initial condition $y(a) = b$. To verify that it also satisfies the differential equation (3), we use the **fundamental theorem of the calculus**: If $f(x)$ is continuous on an interval I, namely $x_1 \leq x \leq x_2$, and

$$F(x) = \int_{x_1}^{x} f(\xi)\, d\xi, \tag{10a}$$

on I, then

$$F'(x) = f(x) \tag{10b}$$

on I. Using this theorem, and chain differentiation [let the $-\int_a^x p(\xi)\, d\xi$ in the exponent be u, say, and write $de^u/dx = (de^u/du)(du/dx)$], it is shown that (9) does indeed satisfy the differential equation $y' + p(x)y = 0$ on an interval I if $p(x)$ is continuous on I.

EXAMPLE 1. Consider the differential equation

$$y' + 2xy = 0 \tag{11}$$

on $-\infty < x < \infty$, over which interval $p(x) = 2x$ is continuous, as we have assumed. Then (7) gives

$$y(x) = Ae^{-\int 2x\, dx} = Ae^{-x^2} \tag{12}$$

on $-\infty < x < \infty$. The graphs of the solutions, (12), are called the solution curves or integral curves corresponding to the given differential equation (11), and these are displayed for several values of A in Fig. 1. Those above and below the x axis correspond to $A > 0$ ($A = 1, 2, 3$) and $A < 0$ ($A = -1, -2$), respectively, and $A = 0$ gives the solution curve $y(x) = 0$. ▮

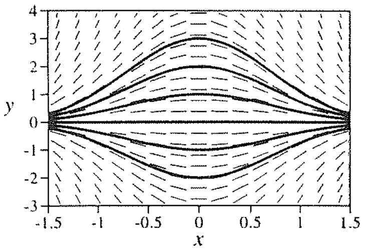


Figure 1. The solution curves and direction field for $y' + 2xy = 0$.

In Example 1 we used the term "solution curve." A **solution curve**, or **integral curve**, corresponding to a differential equation $y'(x) = f(x, y)$, is simply the graph of a solution to that equation.

Besides several solution curves, Fig. 1 also contains a field of lineal elements through a discrete set of points, or grid. By a **lineal element** at a point (x_0, y_0), corresponding to a differential equation $y'(x) = f(x, y)$, we mean a short straight-line segment through the point (x_0, y_0), centered at that point and having the slope $f(x_0, y_0)$. That is, each lineal element has the same slope as the solution curve passing through that point and is therefore a small part of the tangent line to that solution curve. The set of all of the lineal elements is called the **direction field** corresponding to the given differential equation.

In intuitive language, the direction field shows the "flow" of solution curves. Given a sufficiently dense (computer) plot of the direction field, one can visualize the various solution curves, or sketch the solution curve through a given initial point.

EXAMPLE 2. Consider the problem

$$(x + 2)y' - xy = 0, \tag{13a}$$

$$y(0) = 3. \tag{13b}$$

Since an initial value is prescribed, let us use (9) rather than (7), with $p(x) = -x/(x+2)$:

$$\begin{aligned} y(x) &= 3e^{\int_0^x \xi\, d\xi/(\xi+2)} = 3e^{[\xi+2-2\ln|\xi+2|]}\Big|_0^x \\ &= 3e^{[x+2-2\ln|x+2|-(2-2\ln 2)]} = 3e^x e^{\ln|x+2|^{-2}} e^{\ln 2^2} \\ &= 12\frac{e^x}{|x+2|^2} = 12\frac{e^x}{x^2+4x+4}, \end{aligned} \tag{14}$$

where we have used the identity $e^{\ln f} = f$. Of course, we could have used (7) instead, and then imposed the initial condition on that result:

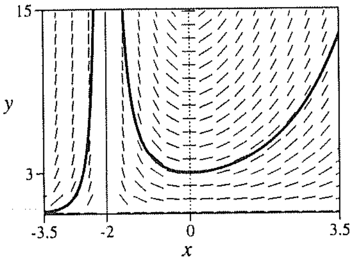


Figure 2. Solution to $(x+2)y' - xy = 0$; $y(0) = 3$, together with the direction field.

$$y(x) = Ae^{\int x\,dx/(x+2)} = Ae^{x+2-2\ln|x+2|} = A\frac{e^2 e^x}{|x+2|^2},$$

where $y(0) = 3 = Ae^2/4$ gives $A = 12e^{-2}$ and, once again, we obtain the solution

$$y(x) = 12\frac{e^x}{x^2+4x+4}. \tag{15}$$

The graph of that solution is given in Fig. 2, in which we show the direction field of the equation (13a), as well.

On what x interval is the solution (15) valid? Recall that the solution (9) was guaranteed over the broadest interval, containing the initial point $x = a$, over which $p(x)$ is continuous. In this case $a = 0$ and $p(x) = -x/(x+2)$, which is undefined (infinite) at $x = -2$, so (15) is valid at least on $-2 < x < \infty$. In fact, the interval of validity cannot be extended any further to the left in this example, because whereas we need both y and y' to have uniquely defined finite values satisfying $(x+2)y' - xy = 0$, both y and y' given by (15) are undefined at $x = -2$. Thus, the interval of validity of (15) is $-2 < x < \infty$. ▮

With the homogeneous problem (3) solved, we now turn to the more difficult case, the nonhomogeneous equation (2). We will show how to solve (2) by two different methods: first, by using an integrating factor and, second, by the method of variation of parameters.

2.2.2. Integrating factor method. To solve (2) by the integrating factor method, we begin by multiplying both sides by a yet to be determined function $\sigma(x)$, so that (2) becomes

$$\sigma y' + \sigma p y = \sigma q. \tag{16}$$

[For (2) and (16) to be equivalent, we require $\sigma(x)$ to be nonzero on the x interval of interest, for if $\sigma(x) = 0$ at one or more points, then it does not follow from (16) that $y' + py$ equals q at those points.] The idea is to seek $\sigma(x)$ so that the left-hand side of (16) is a derivative:

$$\sigma y' + \sigma p y = \frac{d}{dx}(\sigma y), \tag{17}$$

because then (16) becomes

$$\frac{d}{dx}(\sigma y) = \sigma q, \tag{18}$$

which can be solved by integration. For instance, to solve the equation $y' + \frac{1}{x}y = 4$, observe that *if* we multiply through by x, then we have $xy' + y = 4x$, or $(xy)' = 4x$, which can be integrated to give $xy = 2x^2 + C$ and hence the solution $y = 2x + C/x$. In this case $\sigma(x)$ was simply x. Such a function $\sigma(x)$ is called an **integrating factor**, and its use is called the integrating factor method. The idea was invented by *Leonhard Euler* (1707–1783), one of the greatest and most prolific mathematicians of all time. He contributed to virtually every branch of mathematics and to the application of mathematics to the science of mechanics.

One might say that the integrating factor method is similar to the familiar method of solving a quadratic equation by completing the square. In completing the square, we add a suitable quantity to both sides so that the left-hand side becomes a "perfect square," and the equation can then be solved by taking square roots. In the integrating factor method we multiply both sides by a suitable quantity so that the left-hand side becomes a "perfect derivative," and the equation can then be solved by integration.

How do we find $\sigma(x)$, if indeed such a function exists? Writing out the right-hand side of (17) gives

$$\sigma y' + \sigma p y = \sigma y' + \sigma' y,$$

which is satisfied identically if we choose $\sigma(x)$ so that

$$\sigma'(x) = p(x)\sigma(x). \tag{19}$$

But (19) is of the same form as (3), with y changed to σ and p to $-p$, so its solution follows from (7) as $\sigma(x) = Ae^{+\int p(x)\,dx}$. From (16), we see that any constant scale factor in σ is inconsequential, so we can take $A = 1$ without loss. Thus, the desired integrating factor is

$$\sigma(x) = e^{\int p(x)\,dx}. \tag{20}$$

Putting this $\sigma(x)$ into (18) gives

$$\frac{d}{dx}\left(e^{\int p(x)\,dx}y\right) = e^{\int p(x)\,dx}q(x),$$

so

$$e^{\int p(x)\,dx}y = \int e^{\int p(x)\,dx}q(x)\,dx + C,$$

or

$$\boxed{y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx}q(x)\,dx + C\right),} \tag{21}$$

where C is an arbitrary constant of integration.

Not only does (21) satisfy (2), but it can be seen from the steps leading to (21) that *every* solution to (2) must be of the form (21). Thus, we call (21) the **general solution** of (2).

EXAMPLE 3. Solve

$$y' + 3y = x. \tag{22}$$

With $p(x) = 3$ and $q(x) = x$, we have

$$e^{\int p(x)\,dx} = e^{\int 3\,dx} = e^{3x},$$

so (21) gives

$$y(x) = e^{-3x}\left(\int e^{3x} x\,dx + C\right) = \frac{x}{3} - \frac{1}{9} + Ce^{-3x}. \tag{23}$$

as the general solution to (22). ▮

If we wish to solve for C, in (21), so as to satisfy an initial condition $y(a) = b$, then it is convenient to return to (20) and use the definite integral $\int_a^x p(\xi)\,d\xi$ in place of the indefinite integral $\int p(x)\,dx$. These integrals differ by at most an arbitrary additive constant, say B, and the scale factor e^B that results, in the right-hand side of (20), can be discarded without loss. Thus, let us change both p integrals in (21) to definite integrals, with a as lower limit. Likewise, the q integral in (21) can be changed to a definite integral with a as lower limit since this step changes the integral by an arbitrary additive constant at most, which constant can be absorbed by the arbitrary constant C. Thus, equivalent to (21), we have

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + C\right),$$

where ζ is zeta. If we impose on this result the initial condition $y(a) = b$, we obtain $y(a) = b = e^{-0}(0 + C)$, where each ξ integral is zero because its lower and upper integration limits are the same. Thus, $C = b$, and

$$\boxed{y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right).} \tag{24}$$

As a partial check, notice that (24) does reduce to (9) in the event that $q(x) = 0$.

Whereas (21) was the general solution to (2), we call (24) a **particular solution** since it corresponds to one particular solution curve, the solution curve through the point (a, b).

EXAMPLE 4. Solve

$$y' - 2xy = \sin x, \tag{25a}$$

$$y(0) = 3. \tag{25b}$$

This time an initial condition is prescribed, so it is more convenient to use (24) than (21). With $p(x) = -2x$, $q(x) = \sin x$, $a = 0$, and $b = 3$, we have

$$e^{\int_0^x p(\xi)\,d\xi} = e^{\int_0^x (-2\xi)\,d\xi} = e^{-x^2},$$

so that (24) gives the desired solution as

$$y(x) = e^{x^2}\left(\int_0^x e^{-\xi^2}\sin\xi\,d\xi + 3\right). \tag{26}$$

The integral in (26) is said to be **nonelementary** in that it cannot be evaluated in closed form in terms of the so-called **elementary functions**: powers of x, trigonometric and inverse trigonometric functions, exponentials, and logarithms. Thus, we will leave it as it is. It can be evaluated in terms of nonelementary functions, or integrated numerically, but such discussion is beyond the scope of this example. ▮

2.2.3. Existence and uniqueness for the linear equation. A fundamental issue, in the theory of differential equations, is whether a given differential equation $F(x, y, y') = 0$ *has* a solution through a given initial point $y(a) = b$ in the x, y plane. That is the question of existence. If a solution does exist, then our next question is: is that solution unique, or is there more than one such solution? That is the question of uniqueness. Finally, if we do indeed have a unique solution, then over what x interval does it apply?

For the linear problem

$$y' + p(x)y = q(x), \qquad y(a) = b, \tag{27}$$

all of these questions are readily answered. Our proof of existence is said to be constructive because we actually *found* a solution, (24). We are also assured that that solution is unique because our derivation of (24) did not offer any alternatives: each step was uniquely implied by the preceding one, as you can verify by reviewing those steps. Nonetheless, let us show how to prove uniqueness, for (27), by a different line of approach.

Suppose that we have two solutions of (27), $y_1(x)$ and $y_2(x)$, on any x interval I containing the initial point a. That is,

$$y_1' + p(x)y_1 = q(x), \qquad y_1(a) = b, \tag{28a}$$

and

$$y_2' + p(x)y_2 = q(x), \qquad y_2(a) = b. \tag{28b}$$

Next, denote the difference $y_1(x) - y_2(x)$ as $u(x)$, say. If we subtract (28b) from (28a), and use the fact that $(f - g)' = f' - g'$, known from the calculus, we obtain the "homogenized" problem

$$u' + p(x)u = 0, \qquad u(a) = 0, \tag{29}$$

on $u(x)$. But $u' + pu = 0$ implies that $\int du/u + \int p\,dx = 0$, which implies that $\ln|u| = -\int p\,dx + C$, which implies that $|u| = \exp\left(-\int p\,dx + C\right)$ and, finally, that $u(x) = A\exp\left(-\int p\,dx\right)$, where A is an arbitrary constant. Since $u(a) = 0$ and the exponential is nonzero, it follows that A must be zero, so $u(x) = y_1(x) - y_2(x)$ must be identically zero. Thus, $y_1(x)$ and $y_2(x)$ must be identical on I. Since $y_1(x)$ and $y_2(x)$ are *any* two solutions of (27), the solution must be unique.

That approach, in proving uniqueness, is somewhat standard. Namely, we suppose that we have two solutions to the given problem, we let their difference be u, say, obtain a homogenized problem on u, and show that u must be identically zero.

Finally, what is the interval of existence of our solution (24)? The only possible breakdown of (24) is that one or more of the integrals might not exist (be convergent). But since p is continuous, by assumption, it follows that the ξ and ζ integrals of p both exist and, indeed, are continuous functions of x and ξ, respectively. Thus $\exp\left(\int_a^\xi p(\zeta)\,d\zeta\right)$ is continuous too, and since q is also continuous by assumption, then the integral of the exponential times q must exist.

In summary, we have the following result.

THEOREM 2.2.1 *Existence and Uniqueness, for the Linear Equation.*
The linear equation $y' + p(x)y = q(x)$ does admit a solution through an initial point $y(a) = b$ if $p(x)$ and $q(x)$ are continuous at a. That solution is unique, and it exists at least on the largest x interval containing $x = a$, over which $p(x)$ and $q(x)$ are continuous.

In Example 1, for instance, $p(x) = 2x$ and $q(x) = 0$ were continuous for all x, so every solution was valid over $-\infty < x < \infty$. In Example 2, $p(x) = -x/(x+2)$ and $q(x) = 0$, so the broadest interval containing the initial point $x = 0$, over which p and q were continuous, was $-2 < x < \infty$ and, sure enough, we found that the solution (15) was valid over that interval but not beyond it because of the singularity in the solution at $x = -2$. We might think of the solution as "inheriting" that singularity from the singular behavior of $p(x) = -x/(x+2)$ at that point.

EXAMPLE 5. *A Narrow Escape.* The condition of continuity of $p(x)$ and $q(x)$ is sufficient to imply the conclusions stated in the theorem, but is not necessary, as illustrated by the problem

$$xy' + 3y = 6x^3. \tag{30}$$

The general solution of (30) is readily found, from (21), to be

$$y(x) = x^3 + \frac{C}{x^3}. \tag{31}$$

The graphs of the solution for several values of C (i.e., the solution curves) are shown in Fig. 3. Now, $p(x) = 3/x$ is continuous for all x except $x = 0$, and $q(x) = 6x^2$ is continuous for all x, so if we append to (30) an initial condition $y(a) = b$, for some $a > 0$, then Theorem 2.2.1 tells us that the solution (31) that passes through that initial point will be valid over the interval $0 < x < \infty$ at least. For instance, if $a = 1$ and $b = 2.5$ (the point P_1), then $C = 1$, and the solution (31) is valid only over $0 < x < \infty$ since it is undefined at $x = 0$ because of the $1/x^3$ term, as can also be seen from the figure. However, if $a = 1$ and $b = 1$ (the point P_2) then $C = 0$, and the solution (31) is valid over the broader interval $-\infty < x < \infty$ because $C = 0$ removes the singular $1/x^3$ term in (31). That is, if the initial point happens to lie on the solution curve $y(x) = x^3$ through the origin, then the solution $y(x) = x^3$ is valid on $-\infty < x < \infty$; if not, then the solution (30) is valid on $0 < x < \infty$ if $a > 0$ and on $-\infty < x < 0$ if $a < 0$. ▮

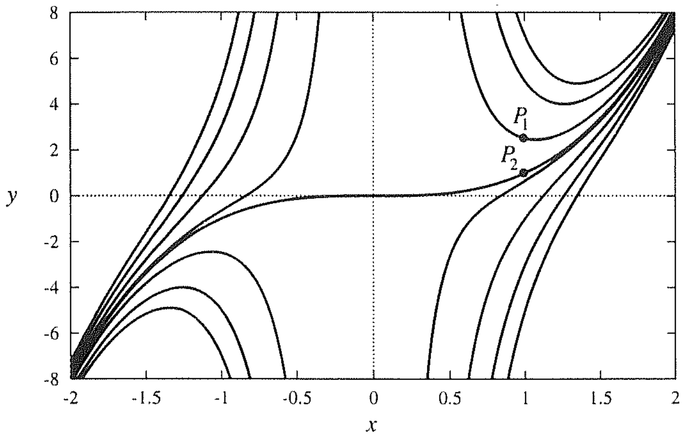


Figure 3. Representative integral curves (31) for the equation (30).

2.2.4. Variation of parameter method.

A second method for the solution of the general first-order linear equation

$$y' + p(x)y = q(x) \tag{32}$$

is the method of **variation of parameters**, due to the great French mathematician *Joseph Louis Lagrange* (1736–1813) who, like Euler, also worked on the applications of mathematics to mechanics, especially celestial mechanics.

Lagrange's method is as follows. We begin by considering the homogeneous version of (32),

$$y' + p(x)y = 0, \tag{33}$$

which is more readily solved. Recall that we solved it by integrating

$$\int \frac{dy}{y} + \int p(x)\,dx = 0$$

and obtaining the general solution

$$y_h(x) = Ae^{-\int p(x)\,dx}. \tag{34}$$

We use the subscript h because $y_h(x)$ is called the **homogeneous solution** of (32). That is, it is the solution of the homogeneous version (33) of the original nonhomogeneous equation (32). [In place of $y_h(x)$, some authors write $y_c(x)$ and call it the **complementary solution**.] Lagrange's idea is to try varying the "parameter" A, the arbitrary constant in (34). Thus, we seek a solution $y(x)$ of the *non*homogeneous equation in the form

$$y(x) = A(x)e^{-\int p(x)\,dx}. \tag{35}$$

(The general idea of seeking a solution of a differential equation in a particular form is important and is developed further in subsequent chapters.)

Putting (35) into (32) gives

$$\left(A'e^{-\int p\,dx} + A(-p)e^{-\int p\,dx}\right) + pAe^{-\int p\,dx} = q. \tag{36}$$

Cancelling the two A terms and solving for A' gives

$$A'(x) = q(x)e^{\int p(x)\,dx}, \tag{37}$$

which can be integrated to give

$$A(x) = \int e^{\int p(x)\,dx}q(x)\,dx + C$$

and hence the general solution

$$y(x) = A(x)e^{-\int p(x)\,dx} = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx}q(x)\,dx + C\right), \tag{38}$$

which is identical to our previous result (21).

It is easy to miss how remarkable is the idea behind Lagrange's method because it starts out looking like a foolish idea and ends up working beautifully. Why do we say that it looks foolish? Because it is completely nonspecific. To explain what we mean by that, let us put Lagrange's idea aside, for a moment, and consider the second-order linear equation $y''+y'-2y = 0$, for instance. In Chapter 3 we will learn to seek solutions of such equations in the exponential form $y = e^{\lambda x}$, where λ is a constant that needs to be determined. Putting that form into $y'' + y' - 2y = 0$ gives the equation $(\lambda^2 + \lambda - 2)e^{\lambda x} = 0$, which implies that λ needs to satisfy the quadratic equation $\lambda^2 + \lambda - 2 = 0$, with roots $\lambda = 1$ and $\lambda = -2$. Thus, we are successful, in this example, in finding two solutions of the assumed form, $y(x) = e^x$ and $y(x) = e^{-2x}$. Notice how easily this idea works. It is easily implemented because most of the work has been done by deciding to look for solutions in the correct form, exponentials, rather than looking within the set of all possible functions. Similarly, if we lose our eyeglasses, the task of finding them is much easier if we know that they are somewhere on our desk, than if we know that they are somewhere in the universe.

Returning to Lagrange's idea, observe that the form (35) is completely nonspecific. That is, *every* function can be expressed in that form by a suitable choice of $A(x)$. Thus, (35) seems useless in that it does not narrow down the search in the least. That's why we say that at first glance Lagrange's idea looks like a foolish idea.

Next, why do we say that, nevertheless, it works beautifully? Notice that the equation (36), governing $A(x)$, is itself a first-order nonhomogeneous equation of the same form as the original equation (32), and looks even harder than (32)–except for the fact that the two A terms cancel, so that we obtain the simple equation (37) that can be solved by direct integration. The cancellation of the two A terms was not serendipitous. For suppose that $A(x)$ is a constant. Then the A' term in (36) drops out, and the two A terms must cancel to zero because if A is a constant then (35) is a solution of the homogeneous equation!

In Chapter 3 we generalize Lagrange's method to higher-order differential equations.

Closure. In this chapter we begin to solve differential equations. In particular, we consider the general first-order linear equation

$$y' + p(x)y = q(x), \tag{39}$$

where $p(x)$ and $q(x)$ are given. We begin with the homogeneous case,

$$y' + p(x)y = 0 \tag{40}$$

because it is simpler, and find its general solution

$$y(x) = Ae^{-\int p(x)\,dx}. \tag{41}$$

If an initial condition $y(a) = b$ is appended to (40), then (41) gives the particular solution of (40) through the initial point (a, b) as

$$y(x) = be^{-\int_a^x p(\xi)\,d\xi}. \tag{42}$$

Turning next to the full nonhomogeneous equation (39), we derive the general solution

$$y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right), \tag{43}$$

first by the integrating factor method, and then again by the method of variation of parameters. Both of these methods will come up again in subsequent sections and chapters.

If an initial condition $y(a) = b$ is appended to (39), then (43) gives the particular solution of (39) through the initial point (a, b) as

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right), \tag{44}$$

which solution is unique, and which is valid on the broadest x interval containing $x = a$, on which both $p(x)$ and $q(x)$ are continuous, and possibly even on a broader interval than that.

Finally, we introduce the idea of lineal elements and the direction field of a differential equation $y' = f(x, y)$.

It is noteworthy that we are successful in finding the general solution of the general first-order linear equation $y' + p(x)y = q(x)$ explicitly and in closed form. For other equation types we may not be so successful, as we shall see.

In closing, we call attention to the exercises, to follow, that introduce additional important special cases, the **Bernoulli, Riccati, d'Alembert-Lagrange**, and **Clairaut equations**. In subsequent sections we occasionally refer to those equations and to those exercises.

Computer software. There are now several powerful computer-algebra systems, such as *Mathematica*, *MATLAB*, and *Maple*, that can be used to implement much of the mathematics presented in this text – numerically, symbolically, and graphically. Consider the application of *Maple*, as a representative software, to the material in this section.

There are two types of applications involved in this section. One entails finding the general solution of a given first-order differential equation, or a particular solution satisfying a given initial condition. These can be carried out on *Maple* using the dsolve command ("function," in *Maple* terminology).

For example, to solve the equation $(x+2)y' - xy = 0$ of Example 2, for $y(x)$, enter

$$\text{dsolve}((x+2)*\text{diff}(y(x),x) - x*y(x) = 0, y(x));$$

(including the semicolon) and return; **dsolve** is the differential equation solver, and **diff** is the derivative command for y'. [The command for y'' would be diff$(y(x), x, x)$, and so on for higher derivatives.] The output is the general solution

$$y(x) = \frac{\_C1 \;\exp(x)}{x^2 + 4x + 4}$$

where $\_C1$ is *Maple* notation for an arbitrary constant.

To solve the same equation, but this time with the initial condition $y(0) = 3$, enter

$$\text{dsolve}(\{(x+2)*\text{diff}(y(x),x) - x*y(x) = 0,\; y(0) = 3\},\; y(x));$$

and return. The output is the particular solution

$$y(x) = 12\frac{\exp(x)}{x^2 + 4x + 4}$$

which agrees with our result in Example 2.

The dsolve command can cope with differential equations that contain unspecified parameters or functions. For example, to solve $y' + p(x)y = 0$, where $p(x)$ is not specified, enter

$$\text{dsolve}(\text{diff}(y(x),x) - p(x)*y(x) = 0,\; y(x));$$

and return. The output is the general solution

$$y(x) = \exp\left(\int p(x)\,dx\right) \_C1$$

The second type of application entails generating the graphical display of various solution curves and/or the direction field for a given differential equation. Both of these tasks can be carried out on *Maple* using the **phaseportrait** command. For example, to obtain the plot shown as Fig. 1, enter

with(DEtools):

to access the phaseportrait command; then return and enter

phaseportrait $(-2 * x * y,\ [x, y],\ x = -1.5..1.5,\ \{[0, -2], [0, -1],$
$[0, 0], [0, 1], [0, 2], [0, 3]\},$ arrows = LINE);

and return. The items within the outer parentheses are as follows:

phaseportrait(right-hand side of $y' = f(x, y)$, [variables], xrange,
{initial points}, optional specification to include direction
field lineal elements and choice of their line thickness);

The yrange is set automatically, but it can be specified as an additional optional item if you wish. All items following the { initial points } are optional, so if you want the yrange to be $-1 \leq y \leq 3$, say, then modify the phaseportrait command as follows:

phaseportrait $(-2 * x * y,\ [x, y],\ x = -1.5..1.5,\ \{[0, -2], [0, -1],$
$[0, 0], [0, 1], [0, 2], [0, 3]\},\ y = -1..3,$ arrows = LINE);

To run phaseportrait over and over, one needs to enter "with(DEtools):" only at the beginning of the session.

To obtain a listing of the mathematical functions and operators (or commands) available in *Maple*, enter **?lib** and return. Within that list one would find such commands as dsolve and phaseportrait. To learn how to use a command enter a question mark, then the command name, then return. For example, type **?dsolve** and return.

In the exercises that follow, and those in subsequent sections, problems are included that require the use of a computer-algebra system such as one of the systems mentioned above. These are important, and we strongly urge you to develop skill in the use of at least one such system in parallel with, but not in place of, developing understanding of the underlying mathematics presented in this text.

EXERCISES 2.2

1. Assuming that $p(x)$ and $q(x)$ are continuous, verify by direct substitution

(a) that (9) satisfies (3)   (b) that (21) satisfies (2)

2. In each case find the general solution, both using the "off-the-shelf" formula (21) and then again by actually carrying out the steps of the integrating factor method. That is, find the integrating factor $\sigma(x)$ and then carry out solution steps analogous to those in our derivation of (21). Understand the x interval on which the equation is defined to be the broadest interval on which both $p(x)$ and $q(x)$ are continuous. For example, in part (a) the x interval is $-\infty < x < \infty$, in part (e) it is any interval on which $\tan x$ is continuous (such as $\pi/2 < x < 3\pi/2$), and in part (f) it is either $-\infty < x < 0$ or $0 < x < \infty$ [to ensure the continuity of $p(x) = 2/x$].

(a) $y' - y = 3e^x$ (b) $y' + 4y = 8$
(c) $y' + y = x^2$ (d) $y' = y - \sin 2x$
(e) $y' - (\tan x)y = 6$ (f) $xy' + 2y = x^3$
(g) $xy' - 2y = x^3$ (h) $y' + (\cot x)y = 2\cos x$
(i) $(x-5)(xy' + 3y) = 2$ (j) $\dfrac{dx}{dy} - 6x = e^y$
(k) $y\dfrac{dx}{dy} - y^5 + 3x = 0$ (l) $y^2\dfrac{dx}{dy} + xy - 4y^2 = 1$
(m) $t\dfrac{dx}{dt} - 4t^5 = x$ (n) $\dfrac{dr}{d\theta} + 2(\cot 2\theta)r = 1$

3.(a)–(n) For the equation given in Exercise 2, solve by the method of variation of parameters. That is, first find the homogeneous solution, then vary the parameter, and so on – as we did in (34)–(37) for the general equation (31).

4.(a)–(n) For the equation given in Exercise 2, find the general solution using computer software (such as *Mathematica*, *MATLAB*, or *Maple*). Verify your result by showing that it does satisfy the given differential equation.

5. Solve $xy' + y = 6x^2$ subject to the given initial condition using any method of this section, and state the (broadest) interval of validity of the solution. Also, sketch the graph of the solution, by hand, and label any key values.

(a) $y(1) = 0$ (b) $y(1) = 2$ (c) $y(2) = 2$
(d) $y(-3) = 18$ (e) $y(-3) = -5$ (f) $y(-2) = 8$

6. Solve $xy' + 2y = x + 2$ subject to the given initial condition using any method of this section, and state the (broadest) interval of validity of the solution. Also, sketch the graph of the solution, by hand, and label any key values.

(a) $y(2) = 0$ (b) $y(0) = 1$ (c) $y(-1) = 1$
(d) $y(1) = 1$ (e) $y(-2) = 0$ (f) $y(-3) = 0$

7. Find the general solution using any method of this section. The answer may be left in implicit form, rather than explicit form, if necessary. HINT: Remember that which variable is the independent variable and which is the dependent variable is a matter of viewpoint, and one can change one's viewpoint. In these problems, consider whether it might be better to regard x as a function of y, and recall from the calculus that $dy/dx = 1/(dx/dy)$.

(a) $\dfrac{dy}{dx} = \dfrac{1}{x + 3e^y}$ (b) $\dfrac{dy}{dx} = \dfrac{1}{6x + y^2}$
(c) $(6y^2 - x)\dfrac{dy}{dx} - y = 0$ (d) $(y^2 \sin y + x)\dfrac{dy}{dx} = y$

8. (*Direction fields*) The direction field concept was discussed within Example 1. For the differential equation given, use computer software to plot the direction field over the specified rectangular region in the x, y plane, as well as the integral curve through the specified point P. Also, if you can identify any integral curves exactly, from an inspection of the direction field, then give the equations of those curves, and verify that they do satisfy the given differential equation.

(a) $y' = 2 + (2x - y)^3$ on $|x| \le 4,\ |y| \le 4;\ P = (2, 1)$
(b) $y' = y(y^2 - 4)$ on $|x| \le 4,\ |y| \le 4;\ P = (1, 1)$
(c) $y' = (3 - y^2)^2$ on $|x| \le 2,\ |y| \le 3;\ P = (0, 0)$
(d) $y' + 2y = e^{-x}$ on $|x| \le 3,\ |y| \le 2;\ P = (0, 0.5)$
(e) $y' = x^2/(y^2 - 1)$ on $|x| \le 3,\ |y| \le 3;\ P = (-3, -3)$
(f) $y' + x = y$ on $|x| \le 20,\ 0 \le y \le 20;\ P = (0, 1)$
(g) $y' = e^x y$ on $0 \le x \le 50,\ 0 \le y \le 50;\ P = (0, 10)$
(h) $y' = x \sin y$ on $0 \le x \le 10,\ 0 \le y \le 10;\ P = (2, 2)$

9. (*Bernoulli equation*) The equation

$$\boxed{y' + p(x)y = q(x)y^n,} \tag{9.1}$$

where n is a constant (not necessarily an integer), is called **Bernoulli's equation**, after the Swiss mathematician *Jakob Bernoulli*. Jakob (1654–1705), his brother Johann (1667–1748), and Johann's son Daniel (1700–1782), are the best known of the eight members of the Bernoulli family who were prominent mathematicians and scientists.

(a) Give the general solution of (9.1) for the special cases $n = 0$ and $n = 1$.

(b) If n is neither 0 nor 1, then (9.1) is nonlinear. Nevertheless, show that by transforming the dependent variable from

$y(x)$ to $v(x)$ according to $v = y^{1-n}$ (for $n \neq 0, 1$), (9.1) can be converted to the equation

$$v' + (1-n)p(x)v = (1-n)q(x), \tag{9.2}$$

which is linear and can be solved by the methods developed in this section. This method of solution was discovered by *Gottfried Wilhelm Leibniz* (1646–1716) in 1696.

10. Use the method suggested in Exercise 9(b) to find the general solution to each of the following.

(a) $y' - 4y = 4y^2$ (b) $xy' - 2y = x^3y^2$
(c) $2xyy' + y^2 = 2x$ (d) $\sqrt{y}(3y' + y) = x$
(e) $y' = y^2$ (f) $y' = xy^3$
(g) $y'' = (y')^2$ HINT: First, let $y'(x) = u(x)$.
(h) $y''' + (y'')^2 = 0$ HINT: First, let $y''(x) = u(x)$.

11. (*Riccati equation*) The equation

$$\boxed{y' = p(x)y^2 + q(x)y + r(x)} \tag{11.1}$$

is called **Riccati's equation**, after the Italian mathematician *Jacopo Francesco Riccati* (1676–1754). The Riccati equation is nonlinear if $p(x)$ is not identically zero. Recall from Exercise 9 that the Bernoulli equation can always be reduced to a linear equation by a suitable change of variables. Likewise, for the Riccati equation, provided that any one particular solution can be found.

Let $Y(x)$ be any one particular solution of (11.1), as found by inspection, trial and error, or any other means. [Depending on $p(x)$, $q(x)$, and $r(x)$, finding such a $Y(x)$ may be easy, or it may prove too great a task.] Show that by changing the dependent variable from $y(x)$ to $u(x)$ according to

$$y = Y(x) + \frac{1}{u} \tag{11.2}$$

the Riccati equation (11.1) can be converted to the equation

$$u' + [2p(x)Y(x) + q(x)]\,u = -p(x), \tag{11.3}$$

which is linear and can be solved by the methods developed in this section. This method of solution was discovered by *Leonhard Euler* (1707–1783) in 1760.

12. Use the method suggested in Exercise 11 to find the general solution to each of the following. Nonelementary integrals, such as $\int \exp{(ax^2)}\,dx$, may be left as is.

(a) $y' - 4y = y^2$ HINT: $Y(x) = -4$
(b) $y' = y^2 - xy + 1$ HINT: $Y(x) = x$
(c) $(\cos x)y' = 1 - y^2$ HINT: $Y(x) = \sin x$
(d) $y' = e^{-x}y^2 - y$ HINT: $Y(x) = 2e^x$
(e) $y' = xy^2 + 2x - x^5$ HINT: See if you can find a $Y(x)$ in the form ax^b.
(f) $y' = (1-y)(2-y)$
(g) $y' = y^2 - 4$
(h) $y' = (2-y)y$

13. (*d'Alembert-Lagrange equation*) The first- order *nonlinear* differential equation

$$\boxed{y = xf(p) + g(p)} \tag{13.1}$$

on $y(x)$, where it will be convenient to denote y' as p, and f and g are given functions of p, is known as a **d'Alembert–Lagrange equation** after the French mathematicians *Jean le Rond d'Alembert* (1717 – 1783) and *Joseph-Louis Lagrange* (1736–1813).

(a) Differentiating (13.1) with respect to x, show that

$$p - f(p) = [xf'(p) + g'(p)]\,\frac{dp}{dx}. \tag{13.2}$$

Observe that this nonlinear equation on $p(x)$ can be converted to a linear equation if we interchange the roles of x and p by now regarding x as the independent variable and p as the dependent variable. Thus, obtain from (13.2) the linear equation

$$\frac{dx}{dp} - \frac{f'(p)}{p - f(p)}x = \frac{g'(p)}{p - f(p)} \tag{13.3}$$

on $x(p)$. Since we have divided by $p - f(p)$ we must restrict $f(p)$ so that $f(p) \neq p$. Solving the simpler equation (13.3) for $x(p)$, the solution of (13.1) is thereby obtained in *parametric form*: $x = x(p)$ from solving (13.3), and $y = x(p)f(p) + g(p)$ from (13.1). This result is the key idea of this exercise, and is illustrated in parts (b)–(c). In parts (d)–(k) we consider a more specialized result, namely, for the case where $f(p)$ happens to have a "fixed point."

(b) To illustrate part (a), consider the equation $y = 2xy' + 3y'$ [i.e., where $f(p) = 2p$ and $g(p) = 3p$], and derive a parametric solution as discussed in (a).

(c) To illustrate part (a), consider the equation $y = x(y' + y'^2)$ [i.e., $f(p) = p + p^2$ and $g(p) = 0$], and derive the parametric solution discussed in (a).

(d) Suppose that $f(p)$ has a **fixed point** P_0, that is, such that $f(P_0) = P_0$. [A given function f may have none, one, or any number of fixed points. They are found as the solutions of the equation $f(p) = p$.] Show that (13.1) then has the straight line

$$y = P_0x + g(P_0) \tag{13.4}$$

cases where the integrals that occur in the general solution of (13.3) are too difficult to evaluate.]
(e) Show that $f(p) = 3p^2$ has two fixed points, $p = 0$ and $p = 1/3$, and hence show that the equation $y = 3xp^2 + g(p)$ has straight-line solutions $y = g(0)$ and $y = \frac{1}{3}x + g\left(\frac{1}{3}\right)$ for any given function g.
(f) Determine all particular solutions of the form (13.4), if any, for the equation $y = x\left(y'^2 - 2y' + 2\right) + e^{y'}$.
(g) Same as (f), for $y = xe^{y'} - 5\cos y'$.
(h) Same as (f), for $y = x\left(y'^2 - 2y'\right) + 6y'^3$.
(i) Same as (f), for $y = x\left(y'^3 - 3y'\right) - 2\sin y'$.
(j) Same as (f), for $y - x\left(y'^2 + 3\right) = y'^5$.
(k) Same as (f), for $y + x\left(2y' + 3\right) = e^{-y'}$.

14. (*Clairaut equation*) For the special case $f(p) = p$, the d'Alembert–Lagrange equation (13.1) in the preceding exercise becomes

$$\boxed{y = xp + g(p),} \tag{14.1}$$

which is known as the **Clairaut equation**, after the French mathematician *Alexis Claude Clairaut* (1713–1765). (Recall that p denotes y' here.)

(a) Verify, by direct substitution into (14.1), that (14.1) admits the family of solutions

$$y = Cx + g(C), \tag{14.2}$$

where C is an arbitrary constant.
(b) Recall that (13.3) does not hold if $f(p) = p$, but (13.2) does. Letting $f(p) = p$ in (13.2), derive the family of solutions (14.2), as well as the additional particular solution given parametrically by

$$x = -g'(p), \tag{14.3a}$$

$$y = -pg'(p) + g(p). \tag{14.3b}$$

(c) To illustrate, find the parametric solution (14.3) for the equation $y = xy' - y'^2$. Show that in this example (14.3) can be gotten into the explicit form $y = x^2/4$ by eliminating the parameter p between (14.3a) and (14.3b). Plot, by hand, the family (14.2), for $C = 0, \pm 1/2, \pm 1, \pm 2$, together with the solution $y = x^2/4$. (Observe, from that plot, that the particular solution $y = x^2/4$ forms an "envelope" of the family of straight-line solutions. Such a solution is called a **singular solution** of the differential equation.)
(d) Instead of a hand plot, do a computer plot of $y = x^2/4$ and the family (14.2), for $C = 0, \pm 0.25, \pm 0.5, \pm 0.75, \dots, \pm 3$, on $-8 \le x \le 8$, $-10 \le y \le 12$.

2.3 Applications of the Linear Equation

In this section we consider representative physical applications that are governed by linear first-order equations: electrical circuits, radioactivity, population dynamics, and mixing problems, with additional applications introduced in the exercises.

2.3.1. Electrical circuits. In Section 1.3 we discussed the mathematical modeling of a mechanical oscillator. The relevant physics was Newton's second law of motion, which relates the net force on a body to its resulting motion. Thus, we needed to find sufficiently accurate expressions for the forces contributed by the individual elements within that system – the forces due to the spring, the friction between the block and the table, and the aerodynamic drag.

In the case of electrical circuits the relevant underlying physics, analogous to Newton's second law for mechanical systems, is provided by Kirchhoff's laws. Instead of forces and displacements in a mechanical system comprised of various elements such as masses and springs, we are interested now in voltages and currents in an electrical system comprised of various elements such as resistors, inductors, and capacitors.

First, by a current we mean a flow of charges: the *current* through a given control surface, such as the cross section of a wire, is the charge per unit time crossing that surface. Each electron carries a negative charge of 1.6×10^{-19} *coulomb*, and each proton carries an equal positive charge. Current is measured in *amperes*, with one ampere being a flow of one coulomb per second. By convention, a current is counted as positive in a given direction if it is the flow of positive charge in that direction. While, in general, currents can involve the flow of positive or negative charges, in an electrical circuit the flow is of negative charges, free electrons. Thus, when one speaks of a current of one ampere in a given direction in an electrical circuit one really means the flow of one coulomb per second of negative charges (electrons) in the opposite direction.

Just as heat flows due to a temperature difference, from one point to another, an electric current flows due to a difference in the electric potential, or *voltage*, measured in *volts*.

We will need to know the relationship between the voltage difference across a given circuit element and the corresponding current flow. The circuit elements of interest here are resistors, inductors, and capacitors.

For a **resistor**, the voltage drop $E(t)$, where t is the time (in seconds), is proportional to the current $i(t)$ through it:

$$E(t) = Ri(t), \tag{1}$$

where the constant of proportionality R is called the *resistance* and is measured in *ohms*; (1) is called **Ohm's law**. By a resistor we usually mean an "off-the-shelf" electrical device, often made of carbon, that offers a specified resistance – such as 100 ohms, 500 ohms, and so on. But even the current-carrying wire in a circuit is itself a resistor, with its resistance directly proportional to its length and inversely proportional to its cross-sectional area, though that resistance is probably negligible compared to that of other resistors in the circuit. The standard symbolic representation of a resistor is shown in Fig. 1.

For an **inductor**, the voltage drop is proportional to the time rate of change of current through it:

$$E(t) = L\frac{di(t)}{dt}, \tag{2}$$

where the constant of proportionality L is called the **inductance** and is measured in *henrys*. Physically, most inductors are coils of wire, hence the symbolic representation shown in Fig. 1.

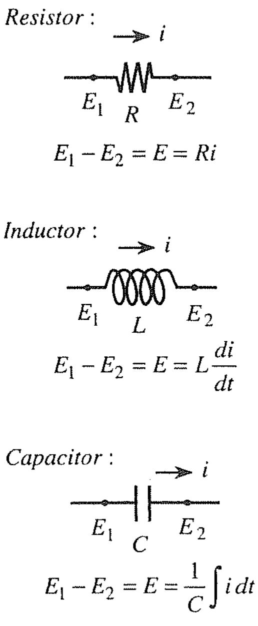

Resistor :
i
E1
R
E2
$E_1 - E_2 = E = Ri$
Inductor :
i
E1
L
E2
$E_1 - E_2 = E = L\frac{di}{dt}$
Capacitor :
i
E1
C
E2
$E_1 - E_2 = E = \frac{1}{C}\int i\,dt$

Figure 1. The circuit elements.

For a **capacitor**, the voltage drop is proportional to the charge $Q(t)$ on the capacitor:

$$E(t) = \frac{1}{C}Q(t), \tag{3}$$

where C is called the **capacitance** and is measured in *farads*. Physically, a capacitor is normally comprised of two plates separated by a gap across which no current flows, and $Q(t)$ is the charge on one plate relative to the other. Though no current flows across the gap, there will be a current $i(t)$ that flows through the circuit that links the two plates and is equal to the time rate of change of charge on the capacitor:

$$i(t) = \frac{dQ(t)}{dt}. \tag{4}$$

From (3) and (4) it follows that the desired voltage/current relation for a capacitor can be expressed as

$$E(t) = \frac{1}{C} \int i(t)\, dt. \tag{5}$$

Now that we have equations (1), (2), and (5) relating the voltage drop to the current, for our various circuit elements, how do we deal with a grouping of such elements within a circuit? The relevant physics that we need, for that purpose, is given by Kirchhoff's laws, named after the German physicist *Gustav Robert Kirchhoff* (1824–1887):

Kirchhoff's current law states that the algebraic sum of the currents approaching (or leaving) any point of a circuit is zero.

Kirchhoff's voltage law states that the algebraic sum of the voltage drops around any loop of a circuit is zero.

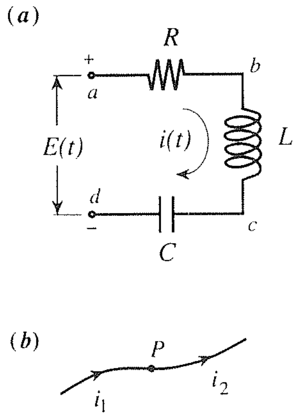


Figure 2. *RLC* circuit.

To apply these ideas, consider the circuit shown in Fig. 2a, consisting of a single loop containing a resistor, an inductor, a capacitor, a voltage source (such as a battery or generator), and the necessary wiring. Let us consider the current $i(t)$ to be positive clockwise; if it actually flows counterclockwise then its numerical value will be negative. In this case Kirchhoff's current law simply says that the current i is a constant from point to point within the circuit and therefore varies only with time. That is, the current law states that at any given point P in the circuit (Fig. 2b), $i_1 + (-i_2) = 0$ or, $i_1 = i_2$. Kirchhoff's voltage law, which is really the self-evident algebraic identity

$$(V_a - V_d) + (V_b - V_a) + (V_c - V_b) + (V_d - V_c) = 0, \tag{6}$$

gives

$$E(t) - Ri - L\frac{di}{dt} - \frac{1}{C}\int i\, dt = 0. \tag{7}$$

The latter is called an **integrodifferential equation** because it contains both derivatives and integrals of the unknown function, but we can convert it to a differential equation in either of two ways.

First, we could differentiate with respect to t to eliminate the integral sign, thereby obtaining

$$\boxed{L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}.} \tag{8}$$

Alternatively, we could use $Q(t)$ instead of $i(t)$ as our dependent variable, for then $\int i\,dt = Q(t)$, and (7) becomes

$$\boxed{L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t).} \tag{9}$$

Either way, we obtain a linear second-order differential equation.

Since we are discussing applications of *first*-order linear equations here, let us treat two special cases.

EXAMPLE 1. *RL Circuit.* If we omit the capacitor from our circuit, then (7) reduces to the first-order linear equation[*]

$$L\frac{di}{dt} + Ri = E(t). \tag{10}$$

If $E(t)$ is a continuous function of time and the current at the initial instant $t = 0$ is $i(0) = i_0$, then the solution to the initial-value problem consisting of (10) plus the initial condition $i(0) = i_0$ is given by (24) in Section 2.2, with "p" $= R/L$ and "q" $= E(t)/L$:

$$i(t) = e^{-\int_0^t \frac{R}{L}\,d\tau}\left(\int_0^t e^{\int_0^\tau \frac{R}{L}\,d\mu}\frac{E(\tau)}{L}\,d\tau + i_0\right),$$

or

$$i(t) = i_0 e^{-Rt/L} + \frac{1}{L}\int_0^t e^{R(\tau - t)/L}E(\tau)\,d\tau \tag{11}$$

over $0 \le t < \infty$, where τ and μ have been used as dummy integration variables.

For instance, if $E(t) = \text{constant} = E_0$, then (11) gives

$$i(t) = i_0 e^{-Rt/L} + \frac{E_0}{R}\left(1 - e^{-Rt/L}\right), \tag{12}$$

or

$$i(t) = \frac{E_0}{R} + \left(i_0 - \frac{E_0}{R}\right)e^{-Rt/L}. \tag{13}$$

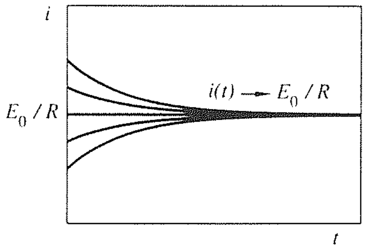


Figure 3. Response $i(t)$ for the case $E(t) = \text{constant} = E_0$; approach to steady state.

As $t \to \infty$, the exponential term in (13) tends to zero, and $i(t) \to E_0/R$. Thus we call the E_0/R term in (13) the **steady-state** solution and the $\left(i_0 - \frac{E_0}{R}\right)e^{-Rt/L}$ term the **transient** part of the solution. The approach to steady state, for several different initial conditions, is shown in Fig. 3.

As another case, let $E(t) = E_0 \sin\omega t$ and $i_0 = 0$. Then (11) gives

$$i(t) = \frac{E_0\omega L}{R^2 + (\omega L)^2}\left(e^{-Rt/L} + \frac{R}{\omega L}\sin\omega t - \cos\omega t\right). \tag{14}$$

[*] It may seem curious that if we try deleting the capacitor by setting $C = 0$, then the capacitor term in (7) becomes infinite rather than zero. Physically, however, one can imagine removing the capacitor, in effect, by moving its plates together until they touch. Since the capacitance C varies as the inverse of the gap dimension, then as the gap diminishes to zero $C \to \infty$, and the capacitor term in the differential equation does indeed drop out because of the $1/C$ factor.

As $t \to \infty$, the exponential term in (14) tends to zero, and we are left with the steady-state solution

$$i(t) \to \frac{E_0 \omega L}{R^2 + (\omega L)^2}\left(\frac{R}{\omega L}\sin\omega t - \cos\omega t\right). \qquad (t \to \infty) \tag{15}$$

Observe that by a steady-state solution we mean that which remains after transients have died out; it is not necessarily a constant. For the case where $i(0) = i_0$ and $E(t) = 0$ the steady-state solution is the constant E_0/R, and for the case where $i(0) = 0$ and $E(t) = E_0 \sin\omega t$ the steady-state solution is the oscillatory function given by (15). ▮

EXAMPLE 2. *RC Circuit.* If, instead of removing the capacitor from the circuit shown in Fig. 2, we remove the inductor (so that $L = 0$), then (8) becomes

$$R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}, \tag{16}$$

which, again, is a first-order linear equation. If we also impose an initial condition $i(0) = i_0$, then

$$i(t) = i_0 e^{-t/RC} + \frac{1}{R}\int_0^t e^{(\tau - t)/RC}\frac{dE(\tau)}{d\tau}\,d\tau \tag{17}$$

gives the solution in terms of i_0 and $E(t)$. ▮

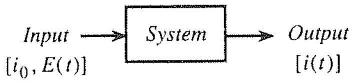


Figure 4. Schematic of the system.

Let us use the electrical circuit problem of Example 1 to make some general remarks. We speak of the initial condition i_0 and the applied voltage $E(t)$ as the **inputs** to the system consisting of the electrical circuit, and the resulting current $i(t)$ as the **output** (or **response**), as denoted symbolically in Fig. 4. From (11), we see that if $i_0 = 0$ and $E(t) = 0$, then $i(t) = 0$: if we put nothing in we get nothing out.*

Consider the inputs and their respective responses separately. If $E(t) = 0$ and $i_0 \neq 0$, then the response

$$i(t) = i_0 e^{-Rt/L}$$

to the input i_0 is seen to be proportional to i_0: if we double i_0 we double its response, if we triple i_0 we triple its response, and so on. Similarly, if $i_0 = 0$ and $E(t)$ is not identically zero, then the response

$$i(t) = \frac{1}{L}\int_0^t e^{R(\tau - t)/L}E(\tau)\,d\tau$$

to the input $E(t)$ is proportional to $E(t)$. This result illustrates an important general property of linear systems: *the response to a particular input is proportional to that input.*

*In contrast with linear initial-value problems, linear *boundary*-value problems can yield nonzero solutions even with zero input – that is, even if the boundary conditions are zero and the equation is homogeneous. These are called **eigensolutions**, and are studied in later chapters.

Further, observe from (11) that the total response $i(t)$ is the sum of the individual responses to i_0 and $E(t)$. This result illustrates the second key property of linear systems: *the response to more than one input is the sum of the responses to the individual inputs.*

In Chapter 3 we prove these two important properties and use them in developing the theory of linear differential equations of second order and higher.

Before closing this discussion of electrical circuits, we wish to emphasize the correspondence, or **analogy**, between the RLC electrical circuit and the mechanical oscillator studied in Section 1.3, and governed by the equation

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t). \tag{18}$$

For we see that both equations (8) (the current formulation) and (9) (the charge formulation) are of exactly the same form as (18). Thus, their mathematical solutions are identical, and hence their physical behavior is identical too. Consider (8), for instance. Comparing it with (18), we note the correspondence

$$L \leftrightarrow m, \quad R \leftrightarrow c, \quad 1/C \leftrightarrow k, \quad i(t) \leftrightarrow x(t), \quad \frac{dE(t)}{dt} \leftrightarrow F(t). \tag{19}$$

Thus, given the values of m, c, k, and the function $F(t)$, we can construct an electrical *analog circuit* by setting $L = m$, $R = c$, $C = 1/k$, and $E(t) = \int F(t)\,dt$. If we also match the initial conditions by setting $i(0) = x(0)$ and $\frac{di}{dt}(0) = \frac{dx}{dt}(0)$, then the resulting current $i(t)$ will be identical to the motion $x(t)$.

Or, we could use (9) to create a different analog, namely,

$$L \leftrightarrow m, \quad R \leftrightarrow c, \quad 1/C \leftrightarrow k, \quad Q(t) \leftrightarrow x(t), \quad E(t) \leftrightarrow F(t). \tag{20}$$

In either case we see that, in mechanical terminology, the inductor provides "inertia" (as does the mass), the resistor provides energy dissipation (as does the friction force), and the capacitor provides a means of energy storage (as does the spring).

Our interest in such analogs is at least twofold. First, to whatever extent we understand the mechanical oscillator, we thereby also understand its electrical analog circuit, and vice versa. Second, if the system is too complex to solve analytically, we may wish to study it experimentally. If so, by virtue of the analogy we have the option of studying whichever is more convenient. For instance, it would no doubt be simpler, experimentally, to study the RLC circuit than the mechanical oscillator.

Finally, just as Hooke's law can be derived theoretically using the governing partial differential equations of the theory of elasticity, our circuit element relations (1)–(5) can be derived using the theory of electromagnetism, the governing equations of which are the celebrated *Maxwell's equations.* We will meet some of the Maxwell's equations later on in this book, when we study scalar and vector field theory.

2.3.2. Radioactive decay; carbon dating. Another important application of first-order linear equations involves radioactive decay and carbon dating.

Radioactive materials, such as carbon–14, einsteinium–253, plutonium–241, radium–226, and thorium–234, are found to decay at a rate that is proportional to the amount of mass present. This observation is consistent with the supposition that the disintegration of a given nucleus, within the mass, is independent of the past or future disintegrations of the other nuclei, for then the number of nuclei disintegrating, per unit time, will be proportional to the total number of nuclei present:

$$\frac{dN}{dt} = -kN, \tag{21}$$

where k is known as the disintegration constant, or decay rate. Actually, the graph of $N(t)$ proceeds in unit steps since $N(t)$ is integer-valued, so $N(t)$ is discontinuous and hence nondifferentiable. However, if N is very large, then the steps are very small compared to N. Thus, we can regard N, approximately, as a continuous function of t and can tolerate the dN/dt derivative in (21). However, it is inconvenient to work with N since one cannot count the number of atoms in a given mass. Thus, we multiply both sides of (21) by the atomic mass, in which case (21) becomes the simple first-order linear equation

$$\frac{dm}{dt} = -km, \tag{22}$$

where $m(t)$ is the total mass, a quantity which is more readily measured. Solving, by means of either (9) or (24) in Section 2.2, we obtain

$$m(t) = m_0 e^{-kt}, \tag{23}$$

where $m(0) = m_0$ is the initial amount of mass (Fig. 5). This result is indeed the exponential decay that is observed experimentally.

Figure 5. Exponential decay.

Since k gives the rate of decay, it can be expressed in terms of the **half-life** T of the material, the time required for any initial amount of mass m_0 to be reduced by half, to $m_0/2$. Then (23) gives

$$\frac{m_0}{2} = m_0 e^{-kT},$$

so $k = (\ln 2)/T$, and (23) can be re-expressed in terms of T as

$$m(t) = m_0 2^{-t/T}. \tag{24}$$

Thus, if $t = T, 2T, 3T, 4T \ldots$, then $m(t) = m_0, m_0/2, m_0/4, m_0/8$, and so on.

Radioactivity has had an important archeological application in connection with **dating**. The basic idea behind any dating technique is to identify a physical process that proceeds at a known rate. If we measure the state of the system now, and we know its state at the initial time, then from these two quantities together with the known rate of the process, we can infer how much time has elapsed; the mathematics enables us to "travel back in time as easily as a wanderer walks up a frozen river."*

*Ivar Ekeland, *Mathematics and the Unexpected*. Chicago: University of Chicago Press, 1988.

In particular, consider carbon dating, developed by the American chemist *Willard Libby* in the 1950's. The essential idea is as follows. Cosmic rays consisting of high-velocity nuclei penetrate the earth's lower atmosphere. Collisions of these nuclei with atmospheric gases produce free neutrons. These, in turn, collide with nitrogen, thus changing some of the nitrogen to carbon-14, which is radioactive, and which decays to nitrogen–14 with a half-life of around 5,570 years. Thus, some of the carbon dioxide which is formed in the atmosphere contains this radioactive C-14. Plants absorb both radioactive and nonradioactive CO_2, and humans and animals inhale both and also eat the plants. Consequently, the countless plants and animals living today contain both C–12 and, to a much lesser extent, its radioactive isotope C-14, in a ratio that is essentially the same from one plant or animal to another.

EXAMPLE 3. *Carbon Dating.* Consider a wood sample that we wish to date. Since C-14 emits approximately 15 beta particles per minute per gram, we can determine how many grams of C-14 are contained in the sample by measuring the rate of beta particle emission. Suppose that we find that the sample contains 0.2 gram of C-14, whereas if it were alive today it would, based upon its weight, contain around 2.6 grams. Thus, we assume that it contained 2.6 grams of C-14 when it died. That mass of C-14 will have decayed, over the subsequent time span t, to 0.2 gram. Then (24) gives

$$0.2 = (2.6)\,2^{-t/5570},$$

and, solving for t, we determine the sample to be around $t = 2{,}100$ years old.

However, it must be emphasized that this method (and the various others that are based upon radioactive decay) depend critically upon assumptions of uniformity. To date the wood sample studied in this example, for instance, we need to know the amount of C-14 present in the sample when the tree died, and what the decay rate was over the time period in question. To apply the method, we assume, first, that the decay rate has remained constant over the time period in question and, second, that the ratio of the amounts of C-14 to C-12 was the same when the tree died as it is today. Observe that although these assumptions are usually stated as fact they can never be proved, since it is too late for direct observation and the only evidence available now is necessarily circumstantial. ∎

2.3.3. Population dynamics. In this application, we are again interested in the variation of a population $N(t)$ with the time t, not the population of atoms this time, but the population of a particular species such as fruit flies or human beings.

According to the simplest model, the rate of change dN/dt is proportional to the population N:

$$\frac{dN}{dt} = \kappa N, \tag{25}$$

where the constant of proportionality κ is the net birth/death rate, that is, the birth rate minus the death rate. As in our discussion of radioactive decay, we regard $N(t)$ as continuous because the unit steps in N are extremely small compared to N itself.

Solving (25), we obtain the exponential behavior

$$N(t) = N_0 e^{\kappa t}, \tag{26}$$

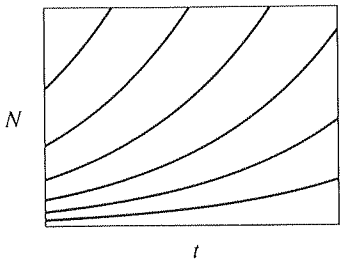


Figure 6. Exponential growth.

where $N(0) = N_0$ is the initial condition. If the death rate exceeds the birth rate, then $\kappa < 0$ and (26) expresses exponential decrease, with $N \to 0$ as $t \to \infty$. That result seems fair enough. However, if $\kappa > 0$, then (26) expresses exponential growth, with $N \to \infty$ as $t \to \infty$, as displayed in Fig. 6 for several different initial conditions N_0. That result is unrealistic because as N becomes sufficiently large other factors will undoubtedly come into play, such as insufficient food or other resources. In other words, we expect that κ will not really be a constant but will vary with N. In particular, we expect it to decrease as N increases. As a simple model of such behavior, suppose that κ varies linearly with N: $\kappa = a - bN$, with a and b positive, so that κ diminishes as N increases, and even becomes negative when N exceeds a/b. Then (25) is to be replaced by the equation

$$\frac{dN}{dt} = (a - bN)N. \tag{27}$$

The latter is known as the **logistic equation**, or the **Verhulst equation**, after the Belgian mathematician *P. F. Verhulst* (1804–1849) who introduced it in his work on population dynamics. Due to the N^2 term, the equation is *nonlinear*, so that the solution that we developed in Section 2.2 does not apply. However, the Verhulst equation is interesting, and we will return to it.

2.3.4. Mixing problems. In this final application we consider a mixing tank with an inflow of $Q(t)$ gallons per minute and an equal outflow, where t is the time; see Fig. 7. The inflow is at a constant concentration c_1 of a particular solute (pounds per gallon), and the tank is constantly stirred, so that the concentration $c(t)$ within the tank is uniform. Hence, the outflow is at concentration $c(t)$. Let v denote the volume within the tank, in gallons; v is a constant because the inflow and outflow rates are equal. To keep track of the instantaneous mass of solute $x(t)$ within the tank, let us carry out a mass balance for the "control volume" V (dashed lines in the figure):

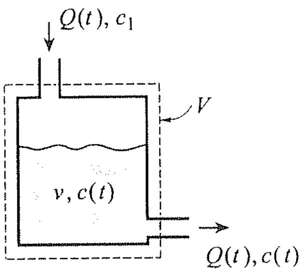


Figure 7. Mixing tank.

$$\begin{array}{c}\text{Rate of increase}\\ \text{of mass of solute}\\ \text{within } V\end{array} = \text{Rate in} - \text{Rate out}, \tag{28}$$

$$\frac{dx}{dt}\frac{\text{lb}}{\text{min}} = \left(Q(t)\frac{\text{gal}}{\text{min}}\right)\left(c_1\frac{\text{lb}}{\text{gal}}\right) - \left(Q(t)\frac{\text{gal}}{\text{min}}\right)\left(c(t)\frac{\text{lb}}{\text{gal}}\right), \tag{29}$$

or, since $c(t) = x(t)/v$,

$$\frac{dx(t)}{dt} + \frac{Q(t)}{v}x(t) = c_1 Q(t), \tag{30}$$

which is a first-order linear equation on $x(t)$. Alternatively, we have the first-order linear equation

$$\frac{dc(t)}{dt} + \frac{Q(t)}{v}c(t) = \frac{c_1 Q(t)}{v} \tag{31}$$

on the concentration $c(t)$.

Recall that in modeling a physical system one needs to incorporate the relevant physics such as Newton's second law or Kirchoff's laws. In the present application, the relevant physics is provided entirely by (28). To better understand (28), suppose we rewrite it with one more term included on the right-hand side:

$$\begin{matrix}\text{Rate of increase} & & \text{Rate} & & \text{Rate} & & \text{Rate of creation} \\ \text{of mass of solute} & = & \text{into} & - & \text{out of} & + & \text{of mass} \\ \text{within } V & & V & & V & & \text{within } V.\end{matrix} \tag{32}$$

The equation (32) is merely a matter of logic, or bookkeeping, not physics. Since (28) follows from (32) only if there is no creation (or destruction) of mass, we can now understand (28) to be a statement of the physical principle of *conservation of mass*, namely, that matter can neither be created nor destroyed (except under exceptional circumstances that are not present in this situation).

Closure. In this section we study applications of first-order linear equations to electrical circuit theory, to radioactivity and population dynamics, and to mixing problems. Although our RLC circuit gives rise to a second-order differential equation, we find that we can work with first-order equations if we omit either the inductor or the capacitor. We will return to the RLC circuit when we discuss second-order equations, so the background provided here, including the expressions for the voltage/current relations and Kirchoff's two laws, will be drawn upon at that time.

The electrical circuit applications also gives us an opportunity to emphasize the extremely important consequences of the linearity of the differential equation upon the relationship between the input and output. The key ideas are that for a linear system: (1) *the response to a particular input is proportional to that input*, and (2) *the response to more than one input is the sum of the responses to the individual inputs*. These ideas are developed and proved in Chapter 3.

EXERCISES 2.3

NOTE: Thus far we have assumed that $p(x)$ and $q(x)$ in $y' + p(x)y = q(x)$ are continuous, yet in applications that may not be the case. In particular, the "input" $q(x)$ may be discontinuous. In Example 1, for instance, $E(t)$ in $L\,di/dt + Ri = E(t)$ may well be discontinuous, such as

$$E(t) = \begin{cases} E_0, & 0 < t < t_1 \\ 0, & t_1 < t < \infty. \end{cases}$$

We state that in such cases, where $E(t)$ has one or more jump discontinuities, the solution (11) [more generally, (24) in Section 2.2] is still valid, and can be used in these exercises.

1. (*RL circuit*) For the RL circuit of Example 1, with $i_0 = 0$ and $E(t) = E_0$, determine the

(a) time required for $i(t)$ to reach 99% of its steady-state value;
(b) resistance R needed to ensure that $i(t)$ will attain 99% of

its steady-state value within 2 seconds, if $L = 0.001$ henry; (c) inductance L needed to ensure that $i(t)$ will attain 99% of its steady-state value within 0.5 seconds, if $R = 50$ ohm.

2. (*RL circuit*) For the RL circuit of Example 1, suppose that $i(0) = i_0$ and that $E(t)$ is as given below. In each case, determine $i(t)$ and identify the steady-state solution. If a steady state does not exist, then state that. Also, sketch the graph of $i(t)$ and label key values.

(a) $$E(t) = \begin{cases} E_0, & 0 < t < t_1 \\ 0, & t_1 < t < \infty \end{cases}$$

(b) $$E(t) = \begin{cases} 0, & 0 < t < t_1 \\ E_0, & t_1 < t < \infty \end{cases}$$

(c) $$E(t) = \begin{cases} 0, & 0 < t < t_1 \\ E_0, & t_1 < t < t_2 \\ 0, & t_2 < t < \infty \end{cases}$$

3. (*RC circuit*) (a) For the RC circuit of Example 2, suppose that $i_0 = 0$ and that $E(t) = E_0 e^{-Rt/L}$. Solve for $i(t)$ and identify the steady-state solution, treating these cases separately: $R^2C \neq L$, and $R^2C = L$. If there does not exist a steady state, then state that. Sketch the graph of $i(t)$ and label any key values.
(b) Same as (a), but with $R = C = 1$ and $E(t) = E_0 \sin t$.

4. Verify that (14) can be re-expressed as

$$i(t) = \frac{E_0 \omega L}{R^2 + (\omega L)^2} e^{-Rt/L} + \frac{E_0}{\sqrt{R^2 + (\omega L)^2}} \sin(\omega t - \phi),$$

where ϕ is the (unique) angle between 0 and $\pi/2$ such that $\tan \phi = \omega L / R$; ϕ is called the **phase angle**.

5. A seashell contains 90% as much C–14 as a living shell of the same size. How old is it? Approximately how many years did it take for its C–14 content to diminish from its initial value to 99% of that?

6. If 10 grams of some radioactive substance will be reduced to 8 grams in 60 years, in how many years will 2 grams be left? In how many years will 0.1 gram be left?

7. If 20% of a radioactive substance disappears in 70 days, what is its half-life?

8. Show that if m_1 and m_2 grams of a radioactive substance are present at times t_1 and t_2, respectively, then its half-life is

$$T = (t_2 - t_1) \frac{\ln 2}{\ln(m_1/m_2)}.$$

9. (*Verhulst equation*) Solve the Verhulst equation (27), subject to the initial condition $N(0) = N_0$, two ways:

(a) by noting that it is a Bernoulli equation;
(b) by noting that it is (also) a Riccati equation.
NOTE: The Bernoulli and Riccati equations, and their solutions, were discussed in the exercises for Section 2.2. (The Verhulst equation can also be solved by the method of separation of variables, which method is the subject of the next section.)

10. (*Mixing tank*) For the mixing tank governed by (31):

(a) Let $Q(t) =$ constant $= Q$ and $c(0) = c_0$. Solve for $c(t)$.
(b) Let $Q(t) = 4$ for $0 < t < 1$ and 2 for $t > 1$, and let $v = c_1 = 1$ and $c(0) = 0$. Solve for $c(t)$. HINT: The application of (24) in Section 2.2 is not so hard when $q(x)$ in the differential equation $y' + p(x)y = q(x)$ is defined piecewise (e.g., as in Exercise 2 above), but is tricky when $p(x)$ is defined piecewise. In this exercise we suggest that you use (24) to solve for $c(t)$ first for $0 < t < 1$, with "a"=0 and "b" = $c(0) = 0$. Then, use that solution to evaluate $c(1)$ and use (24) again, for $1 < t < \infty$, this time with "a"= 1 and "b" = $c(1)$, where $c(1)$ has already been determined.
(c) Let $Q(t) = 2$, $c_1 = 0$, $v = 1$, and $c(0) = 0.3$. Solve for $c(t)$ and thus show that although $c(t) \to 0$ as $t \to \infty$, it never actually reduces to zero, so that it is not possible to wash every molecule of solute out of the tank. Does this result make sense? Explain.

11. (*Mass on an inclined plane*) The equation $mx'' + cx' = mg \sin \alpha$ governs the straight-line displacement $x(t)$ of a mass m along a plane that is inclined at an angle α with respect to the horizontal, if it slides under the action of gravity and friction. If $x(0) = 0$ and $x'(0) = 0$, solve for $x(t)$. HINT: First, integrate the equation once with respect to t to reduce it to a first-order linear equation.

12. (*Free fall; terminal velocity*) The equation of motion of a body of mass m falling vertically under the action of a downward gravitational force mg and an upward aerodynamic drag force $f(v)$, is

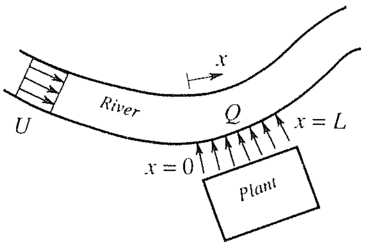
$$mv' = mg - f(v), \tag{12.1}$$

where $v(t)$ is the velocity [so that $v'(t)$ is the acceleration]. The determination of the form of $f(v)$, for the given body shape, would require either careful wind tunnel measurements, or sophisticated theoretical and/or numerical analysis, the result being a plot of the nondimensional drag coefficient versus the nondimensional Reynolds number. All we need to know here is that for a variety of body shapes, the result of such an analysis is the determination that $f(v)$ can be approximated (over some limited range of velocities) in the form cv^{β}, for suitable constants c and β. For low velocities (more precisely, for low Reynolds numbers) $\beta \approx 1$, and for high velocities (i.e., for high Reynolds numbers) $\beta \approx 2$.

(a) Solve (12.1), together with the initial condition $v(0) = 0$, for the case where $f(v) \approx cv$. What is the terminal (i.e., steady-state) velocity?
(b) Same as (a), for $f(v) \approx cv^2$. HINT: Read Exercise 11 in Section 2.2.

13. (*Light extinction*) As light passes through window glass some of it is absorbed. If x is a coordinate normal to the glass (with $x = 0$ at the incident face) and $I(x)$ is the light intensity at x, then the fractional loss in intensity, $-dI/I$ (with the minus sign included because dI will be negative), will be proportional to dx: $-dI/I = k\,dx$, where k is a positive constant. Thus, $I(x)$ satisfies the differential equation $I'(x) = -kI(x)$. The problem: If 80% of the light penetrates a 1-inch thick slab of this glass, how thin must the glass be to let 95% penetrate? NOTE: Your answer should be numerical, not in terms of an unknown k.

14. (*Pollution in a river*) Suppose that a pollutant is discharged into a river at a steady rate Q (grams/second) over a distance L, as sketched in the figure, and we wish to determine the distribution of pollutant in the river – that is, its concentration c (grams/meter^3). Measure x as arc length along the river, positive downstream. The river flows with velocity U (meters/second) and has a cross-sectional area A (meters^2), both of which, for simplicity, we assume to be constant. Also for simplicity, suppose that c is a function of x only. That is, it is a constant over each cross section of the stream. This is evidently a poor approximation near the interval $0 < x < L$, where we expect appreciable across-stream and vertical variations in c, but it should suffice if we are concerned mostly with the far field, that is, more than several river widths upstream or downstream of the interval $0 < x < L$. Then it can be shown that $c(x)$ is governed by the differential equation



$$kc'' - Uc' - \beta c = -\frac{Q(x)}{A}, \quad (-\infty < x < \infty) \qquad (14.1)$$

where k (meters^2/second) is a diffusion constant, β (grams per second per gram) is a chemical decay constant, and $Q(x)$ is the constant Q over $0 < x < L$ and 0 outside that interval. [Physically, (14.1) expresses a mass balance between the *input* $-Q(x)/A$, the transport of pollutant by *diffusion*, kc'', the transport of pollutant by *convection* with the moving stream, Uc', and by disappearance through *chemical decay*, βc.] We assume that the river is clear upstream; that is, we have the initial condition $c(-\infty) = 0$.

(a) Let $L = \infty$. Suppose that k is sufficiently small so that we can neglect the diffusion term. Then (14.1) reduces to the first-order linear equation $Uc' + \beta c = Q(x)/A$. Solve for $c(x)$ and sketch its graph, labeling any key values.
(b) Repeat part (a) for the case where L is finite.

15. (*Newton's law of cooling*) Suppose that a body initially at a uniform temperature u_0 is exposed to a surrounding environment that is at a lower temperature U. Then the outer portion of the body will cool relative to its interior, and this temperature differential within the body will cause heat to flow from the interior to the surface. If the body is a sufficiently good conductor of heat so that the heat transfer within the body is much more rapid than the rate of heat loss to the environment at the outer surface, then it can be assumed, as an approximation, that heat transfer will be so rapid that the interior temperature will adjust to the surface temperature instantaneously, and the body will be at a uniform temperature $u(t)$ at each instant t. **Newton's law of cooling** states that the time rate of change of $u(t)$ will be proportional to the instantaneous temperature difference $U - u$, so that

$$\frac{du}{dt} = k(U - u), \qquad (15.1)$$

where k is a constant.

(a) Solve (15.1) for $u(t)$ subject to the initial condition $u(0) = u_0$. NOTE: Actually, it is not necessary that $U < u_0$; (15.1) is equally valid if $U > u_0$. In most physical applications, however, one is interested in a hot body (such as a cup of coffee or a hot ingot) in a cooler environment.
(b) An interesting application of (15.1) occurs in connection with the determination of the time of death in a homicide. Suppose that a body is discovered at a time T after death and its temperature is measured to be 90° F. We wish to solve for T. Suppose that the ambient temperature is $U = 70°$ F and assume that $u_0 = 98.6°$ F. Putting this information into the solution to (15.1) we can solve for T, provided that we know

k, but we don't. Proceeding indirectly, we can infer the value of k by taking one more temperature reading. Thus, suppose that we wait an hour and again measure the temperature of the body, and find that $u(T+1) = 87°$ F. Use this information to solve for T.

16. (*Compound interest*) Suppose that a sum of money earns interest at a rate k, compounded yearly, monthly, weekly, or even daily. If it is compounded continuously, then $dS/dt = kS$, where $S(t)$ denotes the sum at time t. If $S(0) = S_0$, then the solution is

$$S(t) = S_0 e^{kt}. \tag{16.1}$$

Instead, suppose that interest is compounded yearly. Then after t years

$$S(t) = S_0(1+k)^t,$$

and if the compounding is done n times per year, then

$$S(t) = S_0\left(1+\frac{k}{n}\right)^{nt}. \tag{16.2}$$

(a) Show that if we let $n \to \infty$ in (16.2), then we do recover the continuous compounding result (16.1). HINT: Recall, from the calculus, that

$$\lim_{m\to\infty}\left(1+\frac{1}{m}\right)^m = e.$$

(b) Let $k = 0.05$ (i.e., 5% interest) and compare $S(t)/S_0$ after 1 year ($t = 1$) if interest is compounded yearly, monthly, weekly, daily, and continuously.

2.4 Separable Equations

2.4.1. Separable equations. The general first-order differential equation is of the form

$$F(x, y, y') = 0. \tag{1}$$

If we can solve (1), by algebra, for y', then we can re-express it in the form

$$y' = f(x, y), \tag{2}$$

which form we take as our starting point for this section.

Actually, it is conceivable that we cannot solve (1) for y'. For instance, the equation

$$xy' - y = \sin y' + 4$$

or, equivalently,

$$F(x, y, y') = xy' - y - \sin y' - 4 = 0,$$

cannot be solved, by algebra, for y' in terms of x and y. However, such cases are rarely encountered in applications and will not be considered here. Thus, we assume that the equation can be expressed in the form (2).

If, further, $f(x, y)$ can be expressed as a function of x times a function of y, so that (2) can be written as

$$\boxed{y' = X(x)Y(y),} \tag{3}$$

then we say that the differential equation is **separable**. For instance, $y' = x\ \exp(x+2y)$ is separable because we can factor $x\ \exp(x+2y)$ as $x\ \exp x$ times $\exp(2y)$, but $y' = 3x - y$ is not.

To solve (3), we divide both sides by $Y(y)$ (if $Y(y) \neq 0$) and integrate both sides with respect to x:

$$\int \frac{1}{Y(y)} y' \, dx = \int X(x) \, dx, \tag{4}$$

or, since $y' dx = dy$, from the differential calculus,

$$\boxed{\int \frac{dy}{Y(y)} = \int X(x) \, dx.} \tag{5}$$

We also know from the integral calculus that if $1/Y(y)$ is a continuous function of y (over the relevant y interval) and $X(x)$ is a continuous function of x (over the relevant x interval), then the two integrals in (5) exist, in which case (5) gives the general solution of (2).

EXAMPLE 1. Solve the equation

$$y' = -y^2. \tag{6}$$

Though not linear, (6) is separable. Separating the variables and integrating gives

$$\int \frac{dy}{y^2} = -\int dx, \tag{7}$$

$$-\frac{1}{y} + C_1 = -x + C_2, \tag{8}$$

where C_1 and C_2 are arbitrary. With $C = C_1 - C_2$, we have the general solution

$$y(x) = \frac{1}{x + C}. \tag{9}$$

If we impose an initial condition $y(0) = y_0$ then we can solve for C and obtain the particular solution

$$y(x) = \frac{1}{x + 1/y_0} = \frac{y_0}{1 + y_0 x}, \tag{10}$$

which is plotted in Fig. 1 for the representative values $y_0 = 1$ and $y_0 = 2$. The solution through the initial point $(0, 1)$ exists over $-1 < x < \infty$, the one through (0,2) exists over $-1/2 < x < \infty$. More generally, the one through $(0, y_0)$ exists over $-1/y_0 < x < \infty$ because the denominator in (10) becomes zero at $x = -1/y_0$. We could plot (10) to the left of that point as well, but such extension of the graph would be meaningless because the point $x = -1/y_0$ serves as a "barrier;" y and y' fail to exist there, so the solution cannot be continued beyond that point. ▮

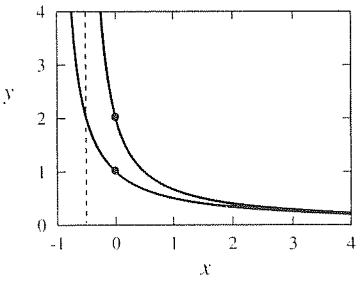


Figure 1. Particular solutions given by (10).

EXAMPLE 2. Solve the initial-value problem

$$y' = \frac{4x}{1 + 2e^y}; \qquad y(0) = 1. \tag{11}$$

Though not linear, the differential equation is separable and can be solved accordingly:

$$\int (1 + 2e^y)\, dy = \int 4x\, dx, \tag{12}$$

$$y + 2e^y = 2x^2 + C. \tag{13}$$

Unfortunately, the latter is a transcendental equation in y, so we cannot solve it for y explicitly as a function of x, as we were able to solve (8). Nevertheless, we can impose the initial condition on (13) to evaluate C: $1 + 2e = 0 + C$, so $C = 1 + 2e$ and the solution is given, in "implicit" form, by

$$y + 2e^y = 2x^2 + 1 + 2e. \tag{14}$$

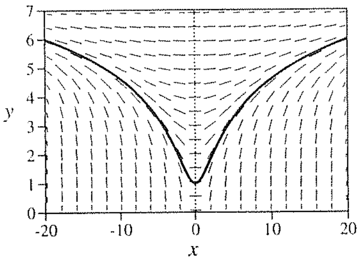


Figure 2. The solution (14) of(11).

The resulting solution is plotted in Fig. 2, along with the direction field. [Actually, we did not plot (14) in Fig. 2; we used the following *Maple* phaseportrait commands to solve (11) and to plot the solution:

```
with (DEtools):
phaseportrait(4 * x/(1 + 2 * exp(y), [x, y], x = −20..20, {[0, 1]}, stepsize = 0.05,
arrows=LINE);
```

where the default stepsize was too large and gave a jagged curve, so we reduced it to 0.05, and where we also included the direction field to give us a feeling for the overall "flow."]

COMMENT 1. Observe that if we use the definite integrals

$$\int_1^y (1 + 2e^y)\, dy = \int_0^x 4x\, dx,$$

with the lower limits dictated by the initial condition $y(0) = 1$, then we bypass the need for an integration constant C and its subsequent evaluation.

COMMENT 2. What is the interval of existence of the solution? In Example 1 we were able to ascertain that interval by direct examination of the solution (10). Here, however, such examination is not possible because the solution (14) is in implicit form. It appears, from Fig. 2, that the solution exists for all x, but of course Fig. 2 covers only $-20 < x < 20$. Equation (14) reveals the asymptotic behavior $2e^y \sim 2x^2$, or $y \sim 2 \ln |x|$ as $|x| \to \infty$, so it seems clear that the solution continues to grow smoothly as $|x|$ increases. ∎

2.4.2. Existence and uniqueness. (Optional) In this section we have begun to solve nonlinear differential equations. Before we get too deeply involved in solution techniques, let us return to the more fundamental questions of existence and uniqueness of solutions. For the linear equation

$$y' + p(x)y = q(x) \tag{15}$$

we have Theorem 2.2.1, which tells us that (15) does admit a solution through an initial point $y(a) = b$ if $p(x)$ and $q(x)$ are continuous at a. That solution is unique, and it exists at least on the largest x interval containing $x = a$, over which $p(x)$ and $q(x)$ are continuous. What can be said about existence and uniqueness for the more general equation $y' = f(x, y)$ (which could, of course, be linear, but, in general, is not)?

THEOREM 2.4.1 *Existence and Uniqueness*
If $f(x, y)$ is continuous on some rectangle R in the x, y plane containing the point (a, b), then the problem

$$y' = f(x, y); \qquad y(a) = b \tag{16}$$

has at least one solution defined on some open x interval* containing $x = a$. If, in addition, $\partial f/\partial y$ is continuous on R, then the solution to (16) is unique on some open interval containing $x = a$.

Notice that whereas Theorem 2.2.1 predicts the minimum interval of existence and uniqueness, Theorem 2.4.1 merely ensures existence and uniqueness over *some* interval; it gives no clue as to how broad that interval will be. Thus, we say that Theorem 2.4.1 is a *local* result; it tells us that under the stipulated conditions all is well locally, in some neighborhood of $x = a$. More informative theorems could be cited, but this one will suffice here.

Let us illustrate Theorem 2.4.1 with two examples.

EXAMPLE 3. The equation

$$y' = \frac{y(y-2)}{x(y-1)} \tag{17}$$

is separable, and separating the variables gives

$$\int \frac{y-1}{y(y-2)}\, dy = \int \frac{dx}{x}. \tag{18}$$

By partial fractions (which method is reviewed in Appendix A),

$$\frac{y-1}{y(y-2)} = \frac{1}{2}\frac{1}{y} + \frac{1}{2}\frac{1}{y-2}. \tag{19}$$

With this result, integration of (18) gives

$$\frac{1}{2}\ln|y| + \frac{1}{2}\ln|y-2| = \ln|x| + C, \qquad (-\infty < C < \infty) \tag{20}$$

where C is the arbitrary constant of integration. Equivalently,

$$\ln\left|\frac{y(y-2)}{x^2}\right| = 2C, \tag{21}$$

*By an **open interval** we mean $x_1 < x < x_2$, and by a **closed interval** we mean $x_1 \leq x \leq x_2$. Thus, a closed interval includes its endpoints, an open interval does not. It is common to use the notation (x_1, x_2) and $[x_1, x_2]$ for open and closed intervals, respectively. Further, $(x_1, x_2]$ means $x_1 < x \leq x_2$, and $[x_1, x_2)$ means $x_1 \leq x < x_2$.

so

$$\left|\frac{y(y-2)}{x^2}\right| = e^{2C} \equiv B, \qquad (0 \le B < \infty) \qquad (22)$$

where B is introduced for convenience and is nonnegative because $\exp(2C)$ is nonnegative. Thus,

$$\frac{y(y-2)}{x^2} = \pm B \equiv A, \qquad (-\infty < A < \infty) \qquad (23)$$

where A replaces the "$\pm B$." Finally, (23) gives $y^2 - 2y - Ax^2 = 0$ so, by the quadratic formula, we have the general solution

$$y(x) = 1 \pm \sqrt{1 + Ax^2} \qquad (24)$$

of (17).

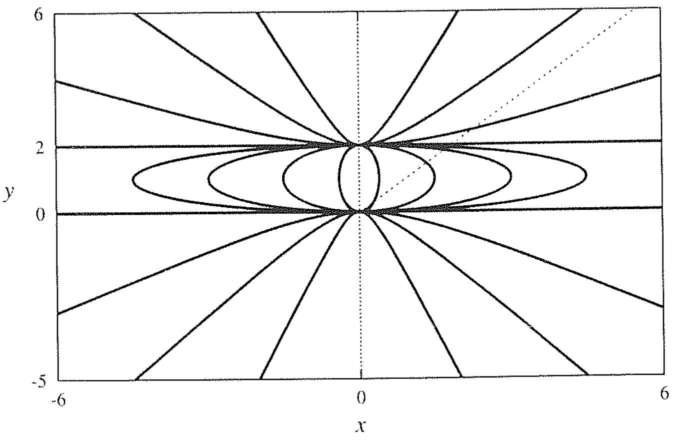


Figure 3. Solution curves corresponding to equation (17).

These solution curves are plotted in Fig. 3. The choice $A = 0$ gives the solution curves $y(x) = 0$ and $y(x) = 2$. As representative of solutions above the line $y = 2$, consider the initial condition $y(1) = 4$. Then (24) gives $y(1) = 4 = 1 \pm \sqrt{1 + A}$, which requires that we select the plus sign and $A = 8$, so $y(x) = 1 + \sqrt{1 + 8x^2}$. As representative of solutions below the line $y = 0$, consider the initial condition $y(1) = -3$. Then (24) gives $y(1) = -3 = 1 \pm \sqrt{1 + A}$, which requires that we select the minus sign and $A = 15$, so $y(x) = 1 - \sqrt{1 + 15x^2}$. Finally, as representative of the solutions between $y = 0$ and $y = 2$, consider the initial condition $y(2) = 3/2$, say. Then $y(2) = 3/2 = 1 \pm \sqrt{1 + 4A}$, so we choose the plus sign and $A = -3/16$, in which case (24) gives $y(x) = 1 + \sqrt{1 - 3x^2/16}$, namely, the upper branch of the ellipse $\frac{3}{16}x^2 + (y-1)^2 = 1$.

In terms of the Existence and Uniqueness Theorem 2.4.1, observe that the conditions of the theorem are met everywhere in the x, y plane except along the vertical line $x = 0$ and the horizontal line $y = 1$, and indeed we do have breakdowns in existence and uniqueness all along these lines. On $x = 0$ (the y axis) there are no solutions through initial points other than $y = 0$ and $y = 2$ (lack of existence), and through each of those points there are an infinite number of solutions (lack of uniqueness). Initial points on the line $y = 1$ are a bit more subtle. We do have elliptical solution curves through each such point, yet at the initial point (on $y = 1$) the slope is infinite, so the differential equation (17) cannot be satisfied at that point. Thus, we have a breakdown in existence for each initial point on $y = 1$. Further, realize that for any such ellipse, between $y = 0$ and $y = 2$, the upper and lower halves are separate solutions. For instance, the ellipse $(3x/4)^2 + (y-1)^2 = 1$, mentioned above, really amounts to the separate solutions $y(x) = 1 \pm \sqrt{1-(3x/4)^2}$, each valid over $-4/3 < x < 4/3$.

COMMENT. Observe that the right side of (17) is asymptotic to y/x as $y \to \pm\infty$, so the solutions of (17) should be asymptotic to the solutions of the simpler equation $y' = y/x$ as $y \to \pm\infty$, namely, the family of straight lines through the origin, and that result can be seen, somewhat, in Fig. 3. ▮

EXAMPLE 4. *Free Fall.* This time consider a physical application. Suppose that a body of mass m is dropped, from rest, at time $t = 0$. With its displacement $x(t)$ measured downward from the point of release, the equation of motion is $mx'' = mg$, where g is the acceleration of gravity and t is the time. Thus,

$$x'' = g, \qquad (0 \le t < \infty) \tag{25a}$$
$$x(0) = 0, \tag{25b}$$
$$x'(0) = 0. \tag{25c}$$

Equation (25a) is of second order, whereas this chapter is about first-order equations, but it is readily integrated twice with respect to t. Doing so, and invoking (25b) and (25c) gives the solution

$$x(t) = \frac{g}{2}t^2, \tag{26}$$

which result is probably familiar to you from a first course in physics.

However, instead of multiplying (25a) through by dt and integrating on t, let us multiply it by dx and integrate on x. Then $x''dx = g\,dx$ and since, from the calculus,

$$x''dx = \frac{dx'}{dt}dx = \frac{dx'}{dt}\frac{dx}{dt}dt = x'\frac{dx'}{dt}dt = x'dx', \tag{27}$$

$x''dx = g\,dx$ becomes

$$x'dx' = g\,dx. \tag{28}$$

Integrating (28) gives

$$\frac{1}{2}x'^2 = gx + A, \tag{29}$$

and $x(0) = x'(0) = 0$ imply that $A = 0$. Thus, we have reduced (25) to the *first*-order problem

$$x' = \sqrt{2g}\,x^{1/2}, \qquad (0 \le t < \infty) \tag{30a}$$
$$x(0) = 0, \tag{30b}$$

which shall now be the focus of this example. Equation (30a) is separable and readily solved. The result is the general solution

$$x(t) = \frac{1}{4}\left(\sqrt{2g}\,t + C\right)^2, \tag{31}$$

which is shown, for various values of C, in Fig. 4. Applying (30b) gives $C = 0$, so (31) gives $x(t) = gt^2/2$, in agreement with (26). However, from the figure we can see that although a solution exists over the full t interval of interest ($t \geq 0$), that solution is not unique because other solutions satisfying both (30a) and (30b) are given by the curve $x(t) = 0$ from the origin up to any point Q, followed by the parabola QR. Physically, the solution OQR corresponds to the mass levitating until time Q, then beginning its descent.

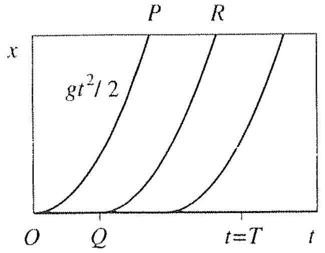


Figure 4. Nonuniqueness of the solution to (30).

Surely that sounds physically impossible, but let us look at the mathematics. We cannot apply Theorem 2.2.1 because (30) is nonlinear, but we can use Theorem 2.4.1 (with x and y replaced by t and x, of course). Since $f(t,x) = \sqrt{2g}\,x^{1/2}$, we see that f is continuous for all $t \geq 0$ and $x \geq 0$, but $f_x(t,x) = \sqrt{g/2}\,x^{-1/2}$ is not continuous over any interval containing the initial point $x = 0$. Thus, the theorem tells us that there does exist a solution over some t interval containing $t = 0$ (which turns out to be the entire positive t axis), but it does not guarantee uniqueness over any such interval, and as it turns out we do not have uniqueness over any such interval.

Next, consider the physics. When we multiply force by distance we get work, and work shows up (in a system without dissipation, as in this example) as energy. Thus, multiplying (25a) by dx and integrating converted the original force equation (Newton's second law) to an energy equation. That is, (29) tells us that the total energy (kinetic plus potential) is conserved; it is constant for all time: $x'^2/2 + (-gx) = $ constant or, equivalently,

$$\frac{1}{2}mx'^2 + (-mgx) = A. \tag{32}$$

Kinetic energy + Potential energy = Constant.

Since $x(0) = x'(0) = 0$, the total energy A is zero. When the mass falls, its kinetic energy becomes positive and its potential energy becomes negative such that their total remains zero for all $t > 0$. However, the energy equation is also satisfied if the released mass levitates for any amount of time and then falls, or if indeed it levitates for all time [that is $x(t) = 0$ for all $t \geq 0$]. Thus, our additional solutions are indeed physically meaningful in that they do satisfy the requirement of conservation of energy. Observe, however, that they do not satisfy the equation of motion (25a) since the insertion of $x(t) = 0$ into that equation gives $0 = g$. Thus, the spurious additional solution $x(t) = 0$ must have entered somewhere between (25) and (30). In fact, we introduced it inadvertently when we multiplied (25a) by dx because $x''dx = g\,dx$ is satisfied not only by $x'' = g$, but also by $dx = 0$ [i.e., by $x(t) = $ constant].

The upshot is that although the solution to (30) is nonunique, a look at our derivation of (30) shows that we should discount the solution $x(t) = 0$ of (30) since it does not also satisfy the original equation of motion $x'' = g$. In that case we are indeed left with the unique solution $x(t) = gt^2/2$, corresponding to the parabola OP in Fig. 4. ▮

It is important to understand that the solution $x(t) = 0$ of (30) is not contained within the general solution (31), for any finite choice of C. Such an additional solution is known as a **singular solution**, and brief consideration of these will be reserved for the exercises.

2.4.3. Applications. Let us study two physical applications of the method of separation of variables.

EXAMPLE 5. *Gravitational Attraction.* **Newton's law of gravitation** states that the force of attraction F exerted by any one point mass M on any other point mass m is*

$$F = G\frac{Mm}{d^2}, \tag{33}$$

where d is the distance between them and $G(= 6.67 \times 10^{-8}\,\text{cm}^3/\text{g sec}^2)$ is called the **universal gravitational constant**; (33) is said to be an *inverse-square law* since the force varies as the inverse square of the distance. (By M and m being point masses, we mean that their sizes are negligible compared with d.)

Consider the linear motion of a rocket of mass m that is launched from the surface of the earth, as sketched in Fig. 5, where M and R are the mass and radius of the earth, respectively. From Newton's second law of motion and his law of gravitation, it follows that the equation of motion of the rocket is

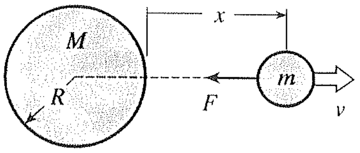


Figure 5. Rocket launch.

$$m\frac{d^2x}{dt^2} = -G\frac{Mm}{(x+R)^2}. \tag{34}$$

Although (34) is a second-order equation, we can reduce it to one of first order by noting that

$$\frac{d^2x}{dt^2} = \frac{d}{dt}\left(\frac{dx}{dt}\right) = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = v\frac{dv}{dx}, \tag{35}$$

*Newton derived (33) from *Kepler's laws* of planetary motion which, in turn, were inferred empirically from the voluminous measurements recorded by the Danish astronomer Tycho Brahe (1546–1601). Usually, in applications (not to mention homework assignments in mechanics), one is given the force exerted on a mass and asked to determine the motion by twice integrating Newton's second law of motion. In deriving (33), however, Newton worked "backwards:" the motion of the planets was supplied in sufficient detail by Kepler's laws, and Newton used those laws to infer the force needed to sustain that motion. It turned out to be an inverse-square force directed toward the sun. Being aware of other such forces between masses, for example, the force that kept his shoes on the floor, Newton then proposed the bold generalization that (33) holds not just between planets and the sun, but between any two bodies in the universe; hence the name *universal law of gravitation*. Just as it is difficult to overestimate the importance of Newton's law of gravitation and its impact upon science, it is also difficult to overestimate how the idea of a force acting at a distance, rather than through physical contact, must have been incredible when first proposed.

In fact, such eminent scientists and mathematicians as Huygens, Leibniz, and John Bernoulli referred to Newton's idea of gravitation as absurd and revolting. Imagine Newton's willingness to stand nonetheless upon the results of his mathematics, in inferring the concept of gravitation, even in the absence of any physical mechanism or physical plausibility, and in the face of such opposition.

Remarkably, *Coulomb's law* subsequently stated an inverse-square type of electrical attraction or repulsion between two charges. Why these two types of force field turn out to be of the same mathematical form is not known. Equally remarkable is the fact that although the forms of the two laws are identical, the magnitudes of the forces are staggeringly different. Specifically, the ratio of the electrical force of repulsion to the gravitational force exerted on each other by two electrons (which is independent of the distance of separation) is 4.17×10^{42}.

where v is the velocity, and where the third equality follows from the chain rule. Thus (34) becomes the first-order equation

$$v\frac{dv}{dx} = -\frac{GM}{(x+R)^2}, \tag{36}$$

which is separable and gives

$$\int v\,dv = -GM\int\frac{dx}{(x+R)^2}, \tag{37}$$

$$\frac{v^2}{2} = \frac{GM}{x+R} + C. \tag{38}$$

If the launch velocity is $v(0) = V$, then (38) gives $C = (V^2/2) - GM/R$, so

$$v = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}} \tag{39}$$

is the desired expression of v as a function of x.

If we wish to know $x(t)$ as well, then we can re-write (39) as

$$\frac{dx}{dt} = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}}, \tag{40}$$

which once again is variable separable and can be solved for $x(t)$. However, let us be content with (39).

Observe from (39) that v decreases monotonically with increasing x, from its initial value $v = V$ to $v = 0$, the latter occurring at

$$x_{max} = \frac{V^2R^2}{2GM - V^2R}. \tag{41}$$

Subsequently, the rocket will be drawn back toward the earth and will strike it with speed V. [We need to choose the negative square root in (39) to track the return motion.] Equation (41) can be simplified by noting that when $x = 0$, the right-hand side of (34) must be $-mg$, where g is the familiar gravitational acceleration at the earth's surface. Thus, $-mg = -GMm/R^2$, so $GM/R^2 = g$, and (41) becomes

$$x_{max} = \frac{V^2R}{2gR - V^2}. \tag{42}$$

We see from (42) that x_{max} increases as V is increased, as one would expect, and becomes infinite as $V \to \sqrt{2gR}$. Thus, the critical value $V_e = \sqrt{2gR}$ is the *escape velocity*. Numerically, $V_e \approx 6.9$ miles/sec.

COMMENT 1. Recall that the law of gravitation (33) applies to two point masses separated by a distance d, whereas the earth is hardly a point mass. Thus, it is appropriate to question the validity of (34). In principle, to find the correct attractive force exerted on the rocket by the earth we need to consider the earth as a collection of point masses dM, compute the force dF induced by each dM, and add the dF's vectorially to find the resultant force F.

This calculation is carried out later, in Section 15.7, and the result, remarkably, is that the resultant F acting at any point P outside the earth (or *any* homogeneous spherical mass), per unit mass at P, is the same as if its entire mass M were concentrated at a single point, namely, at its center! Thus, the earth might as well be thought of as a point mass, of mass M, located at its center, so (34) is exactly true, if we are willing to approximate the earth as a homogeneous sphere.

COMMENT 2. The steps in (35), whereby we were able to reduce our second-order equation (34) to the first-order equation (36), were not limited to this specific application. They apply whenever the force is a function of x alone, for if we apply (35) to the equation

$$m\frac{d^2x}{dt^2} = f(x), \tag{43}$$

we get the separable first-order equation

$$mv\frac{dv}{dx} = f(x) \tag{44}$$

with solution

$$\frac{mv^2}{2} = \int f(x)\,dx + C \tag{45}$$

or, equivalently,

$$\left.\frac{mv^2}{2}\right|_{x_1}^{x_2} = \int_{x_1}^{x_2} f(\xi)\,d\xi. \tag{46}$$

In the language of mechanics, the right-hand side is the work done by the force $f(x)$ as the body moves from x_1 to x_2, and $mv^2/2$ is the kinetic energy. Thus, the physical significance of (46) is that it is a work-energy statement: the change in the kinetic energy of the body is equal to the work done on it.

COMMENT 3. Observe the change in viewpoint as we progressed from (34) to (36). Until the third equality in (35), we regarded x and v as dependent variables – functions of the independent variable t. But beginning with the right-hand side of that equality, we began to regard v as a function of x. However, once we solved (36) for v in terms of x, in (39), we replaced v by dx/dt, and x changed from independent variable to dependent variable once again. In general, then, which variable is regarded as the independent variable and which is the dependent variable is not so much figured out, as it is a decision that we make, and that decision, or viewpoint, can sometimes change, profitably, over the course of the solution. ▮

EXAMPLE 6. *Verhulst Population Model.* Consider the Verhulst population model

$$N'(t) = (a - bN)N; \qquad N(0) = N_0 \tag{47}$$

that was introduced in Section 2.3.3, where $N(t)$ is the population of a given species. This example emphasizes that a given equation might be solvable by a number of different methods. Though (47) is not a linear equation, it is both a Bernoulli equation and a Riccati equation, which equations were discussed in the exercises of Section 2.2. Now we see that it is also separable, since the right side is a function of N [namely, $(a-bN)N$] times a function of t (namely, 1). Thus,

$$\int \frac{dN}{(a-bN)N} = \int dt. \tag{48}$$

By partial fractions,

$$\frac{1}{(a-bN)N} = -\frac{1}{b}\frac{1}{(N-\frac{a}{b})N} = -\frac{1}{a}\frac{1}{N-\frac{a}{b}} + \frac{1}{a}\frac{1}{N}$$

so (48) gives

$$-\frac{1}{a}\ln\left|N-\frac{a}{b}\right| + \frac{1}{a}\ln N = t + C, \tag{49}$$

where C is an arbitrary constant ($-\infty < C < \infty$). [Whether we write $\ln|N|$ or $\ln N$ in (49) is immaterial since $N > 0$.] Equivalently,

$$\left|\frac{N}{N-\frac{a}{b}}\right|^{1/a} = e^{t+C}, \qquad \left|\frac{N}{N-\frac{a}{b}}\right| = e^{at+aC} = Be^{at}, \tag{50}$$

where we have replaced $\exp(aC)$ by B, so $0 \le B < \infty$. Thus

$$\frac{N}{N-a/b} = \pm Be^{at} \equiv Ae^{at}, \tag{51}$$

where A is arbitrary ($-\infty < A < \infty$). Finally, imposing the initial condition $N(0) = N_0$ gives $A = N_0/(N_0 - a/b)$, and putting that expression into (50) and solving for N gives

$$N(t) = \frac{aN_0}{(a-bN_0)e^{-at} + bN_0}. \tag{52}$$

What can be learned of the behavior of $N(t)$ from (52)? We can see from (52) that for every initial value N_0 (other than $N_0 = 0$), $N(t)$ tends to the constant value a/b as $t \to \infty$. [If $N_0 = 0$, then $N(t) = 0$ for all t, as it should, because if a species starts with no members it can hardly wax or wane.] Beyond observing that asymptotic information, it is an excellent idea to plot the results, especially now that one has such powerful and convenient computer software for that purpose. However, observe that the solution (52) contains the three parameters a, b, and N_0, and to use plotting software we need to choose numerical values for these parameters. If, for instance, we wish to plot $N(t)$ versus t for five values of a, five of b, and five of N_0, then we will be generating $5^3 = 125$ curves! Thus, the point is that if we wish to do a parametric study of the solution (i.e., examine the solution for a range of values of the various parameters), then there is a serious problem with managing all of the needed plots. In Section 2.4.4 below, we offer advice on how to deal with this common and serious predicament. ▮

2.4.4. Nondimensionalization. (Optional) One can usually reduce the number of parameters in a problem, sometimes dramatically, by a suitable scaling of the independent and dependent variables so that the new variables are nondimensional (i.e., dimensionless).

EXAMPLE 7. *Example 6, Continued.* To begin such a process of nondimensionalization, we list all dependent and independent variables and parameters, and their dimensions:

	Variable	Dimensions	Parameter	Dimensions
Independent:	t	time	a	$1/\text{time}$
Dependent:	N	number	b	$1/[(\text{time})(\text{number})]$
			N_0	number

(By number we mean the number of living members of the species.) How did we know that a has dimensions of $1/\text{time}$, and that b has dimensions of $1/[(\text{time})(\text{number})]$? From the differential equation $\frac{dN}{dt} = aN - bN^2$. That is, the dimensions of the term on the left are number/time, so the dimensions of aN and bN^2 must be the same. Dimensionally, then, aN = number/time, so $a = 1/\text{time}$. Similarly, bN^2 = number/time, so $b = 1/[(\text{time})(\text{number})]$.

Next, we nondimensionalize the independent and dependent variables (t and N) using suitable combinations of the parameters. From the parameter list, observe that $1/a$ has dimensions of time and can therefore be used as a "reference time" to nondimensionalize the independent variable t. That is, we can introduce a nondimensional version of t, say $\bar{t}$, by $\bar{t} = t/(1/a) = at$.

Next, we need to nondimensionalize the dependent variable N. From the parameter list, observe that N_0 has dimensions of number, so let us introduce a nondimensional version of N, say $\overline{N}$, by $\overline{N} = N/N_0$. In case the notion of nondimensionalization still seems unclear, realize that it is merely a change of variables, from t and N to $\bar{t}$ and $\overline{N}$; a rather simple change of variables in fact, since $\bar{t}$ is simply a constant times t, and $\overline{N}$ is simply a constant times N.

Putting $t = \bar{t}/a$ and $N = N_0\overline{N}$ into (47) gives

$$aN_0\frac{d\overline{N}}{d\bar{t}} = \left(a - bN_0\overline{N}\right)N_0\overline{N}; \qquad N_0\overline{N}(0) = N_0, \tag{53}$$

where the left side of the differential equation follows from the chain differentiation $\frac{dN}{dt} = \frac{dN}{d\overline{N}}\frac{d\overline{N}}{d\bar{t}}\frac{d\bar{t}}{dt} = (N_0)\left(\frac{d\overline{N}}{d\bar{t}}\right)(a) = aN_0\frac{d\overline{N}}{d\bar{t}}$. (More simply, but less rigorously, we could merely replace the dN in dN/dt by $N_0 d\overline{N}$ and the dt by $d\bar{t}/a$.) Simplifying (53) gives

$$\frac{d\overline{N}}{d\bar{t}} = \left(1 - \alpha\overline{N}\right)\overline{N}; \qquad \overline{N}(0) = 1, \tag{54}$$

where $\alpha \equiv bN_0/a$. Thus, (54) contains only the single parameter α. The solution of (54) is

$$\overline{N}(\bar{t}) = \frac{1}{\alpha + (1-\alpha)e^{-\bar{t}}}. \tag{55}$$

The idea is that if we plot $\overline{N}(\overline{t})$ versus $\overline{t}$ (rather than $N(t)$ versus t), then we have only the *one*-parameter family of solutions given by (55), where the parameter is the nondimensional quantity $\alpha = bN_0/a$. Those solutions are shown in Fig. 6 for several different values of α. As $\overline{t} \to \infty$ (and hence $t \to \infty$), $\overline{N} \to 1/\alpha$, so $N/N_0 \to 1/(bN_0/a)$, or $N \to a/b$, as found in Example 4.

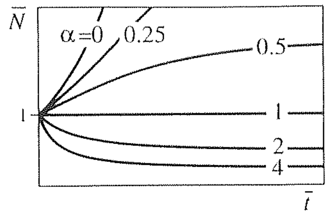


Figure 6. Nondimensional solution of Verhulst problem.

COMMENT. The nondimensionalization of the independent and dependent variables can often be done in more than one way. In the present example, for instance, we used N_0 to nondimensionalize N: $\overline{N} = N/N_0$. However, a/b also has the dimensions of number, so we could have defined $\overline{N} = N/(a/b) = bN/a$ instead. Similarly, we could have nondimensinalized t differently, as $\overline{t} = N_0bt$, because N_0b has dimensions of $1/$time. Any nondimensionalization will work, and we leave these other choices for the exercises. ▮

EXAMPLE 8. *Example 5, Continued.* As one more example of the simplifying use of nondimensionalization, consider the initial-value problem

$$\frac{dx}{dt} = \sqrt{V^2 - \frac{2GM}{R}\frac{x}{x+R}}; \qquad x(0) = R \tag{56}$$

from Example 5. As above, we begin by listing all variables and parameters, and their dimensions:

	Variable	Dimensions	Parameter	Dimensions
Independent:	t	time	V	length/time
Dependent:	x	length	R	length

We didn't bother with G and M in the parameter list because V and R are all we need to nondimensionalize t and x. Specifically, R has dimensions of length, so we can choose $\overline{x} = x/R$, and R/V has dimensions of time, so we can choose $\overline{t} = t/(R/V)$. Putting $x = R\overline{x}$ and $t = R\overline{t}/V$ into (56) gives

$$\frac{R}{R/V}\frac{d\overline{x}}{d\overline{t}} = \sqrt{V^2 - \frac{2GM}{R}\frac{R\overline{x}}{R\overline{x}+R}}; \qquad R\overline{x}(0) = R \tag{57}$$

or

$$\frac{d\overline{x}}{d\overline{t}} = \sqrt{1 - \alpha\frac{\overline{x}}{\overline{x}+1}}; \qquad \overline{x}(0) = 1, \tag{58}$$

with the single parameter $\alpha \equiv 2GM/RV^2$. Since all other quantities in the final differential equation are nondimensional, it follows that α must be nondimensional as well, as could be checked from the known dimensions of G, M, R, and V.

Of course, whereas we've used the generic dimensions "time" and "length," we could have used specific dimensions such as seconds and meters. ▮

It is common in engineering and science to nondimensionalize the governing equations and initial or boundary conditions even before beginning the solution, so as to reduce the number of parameters as much as possible. In each of the foregoing two examples we ended up with a single parameter, but the final number of parameters will vary from case to case. The nondimensional parameters that result (such as α in Example 6) are sometimes well known and of great importance. For instance, if one nondimensionalizes the differential equations governing fluid flow, two nondimensional parameters that arise are the *Reynolds number* Re and the *Mach number* M. Without getting into the fluid mechanics, let it suffice to say that the Reynolds number is a measure of the relative importance of viscous effects to inertial effects: if the Reynolds number is sufficiently large one can neglect the viscous terms in the governing equations of motion, and if it is sufficiently small then one can neglect the inertial terms. Similarly, the Mach number is a measure of the importance of the effects of the compressibility of the fluid: if M is sufficiently small then one can neglect those effects and consider the fluid to be incompressible. In fact, any given approximation that is made in engineering science is probably based upon whether some relevant nondimensional parameter is sufficiently large or small, for one is always neglecting one effect *relative* to others.

Closure. We see that the method of separation of variables is relatively simple: one separates the variables and integrates. Thus, given a specific differential equation, one is well advised to see immediately if the equation is of separable type and, if it is, to solve by separation of variables. Of course, it might turn out that one or both of the integrations are difficult, but the general rule of thumb is that there is a conservation of difficulty, so that if the integrations are difficult, then an equivalent difficulty will show up if one tries a different solution technique.

In the last part of this section we discuss the idea of nondimensionalization. The latter is not central to the topic of this section, separation of variables, but arises tangentially with regard to the efficient management of systems that contain numerous parameters, which situation makes graphical display and general understanding of the results more difficult.

Computer software. A potential difficulty with the method of separation of variables is that the integrations involved may be difficult. Using *Maple*, for instance, integrations can be carried out using the **int** command. For example, to evaluate the integral on the left side of (48), enter

$$\text{int}(1/((a - b*N)*N),\ N);$$

and return. The output is

$$-\frac{\ln(a - bN)}{a} + \frac{\ln(N)}{a}$$

which (to within an additive constant) is the same as the left side of equation (49). That is not to say that all integrals can be evaluated in closed form by computer software. To illustrate the use of the int command for definite integrals, consider the integral of e^{-ax} from $x = 0$ to $x = \infty$. Enter

$$\text{int}(\exp(-x),\ x = 0..\text{infinity});$$

and return. The output is 1.

Of course, if we are going to use *Maple* to evaluate the integrals that arise, then we might as well see if the *Maple* dsolve command will solve the differential equation directly, as is discussed at the end of Section 2.2. For instance, to solve the equation $y' = -3x^2y^5$, enter

$$\text{dsolve}(\text{diff}(y(x), x) = -3 * x\hat{}2 * (y(x))\hat{}5,\ y(x));$$

and return. The output is the general solution

$$\frac{1}{y(x)^4} = 4x^3 + \_C1$$

in implicit form, where $\_C1$ is the arbitrary constant of integration.

EXERCISES 2.4

NOTE: Solutions should be expressed in explicit form if possible.

1. Use separation of variables to find the general solution. Then, obtain the particular solution satisfying the given initial condition. Sketch the graph of the solution, showing the key features, and label any key values.

(a) $y' - 3x^2e^{-y} = 0;\quad y(0) = 0$
(b) $y' = 6x^2 + 5;\quad y(0) = 0$
(c) $y' + 4y = 0;\quad y(-1) = 0$
(d) $y' = 1 + y^2;\quad y(2) = 5$
(e) $y' = (y^2 - y)e^x;\quad y(0) = 2$
(f) $y' = y^2 + y - 6;\quad y(5) = 10$
(g) $y' = y(y + 3);\quad y(0) = -4$
(h) $y' = 6\dfrac{y \ln y}{x};\quad y(1) = e$
(i) $y' = e^{x+2y};\quad y(0) = 1$
(j) $y' = \dfrac{y}{2x};\quad y(3) = -1$
(k) $y' + 3y(y + 1)\sin 2x = 0;\quad y(0) = 1$
(l) $y = \ln y';\quad y(0) = 5$
(m) $y' = y \ln y;\quad y(0) = 5$
(n) $y' + 2y = y^2 + 1;\quad y(-3) = 0$

2.(a)–(n) For the equation given in Exercise 1, use computer software to solve for $y(x)$. Verify, by direct substitution, that your solution does satisfy the given differential equation and initial condition.

3. The problem $du/dt = k(U - u)$; $u(0) = u_0$, where k and U are constants, occurred in the exercises for Section 2.3 in connection with Newton's law of cooling. Solve by separation of variables.

4. The Verhulst population problem

$$N'(t) = (a - bN)N; \qquad N(0) = N_0 \qquad (a > 0,\ b > 0)$$

was studied in Section 2.3 and solved as a Bernoulli equation, and also as a Riccati equation. Here we ask you to solve it by separation of variables.

5. The Bernoulli equation $y' + p(x)y = q(x)y^n$ is not variable separable, in general, but it is if p and q are constants, if one of the functions $p(x)$ and $q(x)$ is zero, or if one is a constant times the other. Obtain the general solution for the case where each is a nonzero constant, for any real number n. HINT: A difficult integral will occur. Our discussion of the Bernoulli equation in the exercises for Section 2.2 should help you to find a change of variables that will simplify that integration.

6. Solve $y' = (6x^2 + 1)/(y - 1)$, subject to the given initial condition.

(a) $y(0) = -2$ (b) $y(0) = 4$ (c) $y(0) = 0$
(d) $y(1) = 3$ (e) $y(2) = 4$ (f) $y(-1) = 0$

7. Solve $y' = (3x^2 - 1)/2y$, subject to the given initial condition.

(a) $y(0) = -3$ (b) $y(0) = -1$ (c) $y(4) = 5$
(d) $y(-1) = 0$ (e) $y(-2) = -4$ (f) $y(1) = -6$

8. (*Homogeneous functions*) A function $f(x_1, \dots, x_n)$ is said to be **homogeneous of degree k** if $f(\lambda x_1, \dots, \lambda x_n) = \lambda^k f(x_1, \dots, x_n)$ for any λ. For example,

$$f(x, y, z) = \frac{4x^5}{y^2 + 3z^2} \sin\left(\frac{y}{z}\right)$$

is homogeneous of degree 3 because

$$f(\lambda x, \lambda y, \lambda z) = \frac{4(\lambda x)^5}{(\lambda y)^2 + 3(\lambda z)^2} \sin\left(\frac{\lambda y}{\lambda z}\right) = \lambda^3 f(x, y, z).$$

State whether f is homogeneous or not. If it is, determine its degree.

(a) $f(x, y) = x^2 + 4y^2 - 7$
(b) $f(x, y, z) = \cos\left(\frac{x - y}{5z}\right)$
(c) $f(x, y) = x^2 - y^2 + 7xz - 3xy$
(d) $f(x, y) = \sin\left(x^2 + y^2\right)$

9. (*Homogeneous equation*) The equation

$$\boxed{y' = f\left(\frac{y}{x}\right)} \tag{9.1}$$

is said to be **homogeneous** because $f(y/x)$ is homogeneous (of degree zero); see the preceding exercise. CAUTION: The term homogeneous is also used to describe a linear differential equation that has zero as its "forcing function" on the right-hand side, as defined in Section 1.2. Thus, one needs to use the context to determine which meaning is intended.

(a) Show, by examples, that (9.1) may, but need not, be separable.
(b) In any case, show that the change of dependent variable $w = y/x$, from $y(x)$ to $w(x)$, reduces (9.1) to the separable form

$$w' = \frac{f(w) - w}{x} \tag{9.1}$$

10. Use the idea contained in the preceding exercise, to find the general solution to each of the following equations.

(a) $y' = \frac{y}{x} + 3\sqrt{\frac{x}{y}}$
(b) $y' = \frac{2y - x}{y - 2x}$
(c) $y' = \frac{xy + 2y^2}{x^2}$
(d) $y' = -\frac{2x + y}{x}$
(e) $y' = e^{y/x} + \frac{y}{x}$

11. (*Almost-homogeneous equation*) (a) Show that

$$y' = \frac{ax + by + c}{dx + ey + f} \qquad (a, b, \ldots, f \text{ constants}) \tag{11.1}$$

can be reduced to homogeneous form by the change of variables $x = u + h$, $y = v + k$, where h and k are suitably chosen constants, provided that $ae - bd \neq 0$.
(b) Thus, find the general solution of $y' = (2x - y - 6)/(x - y - 3)$.
(c) Similarly, for $y' = (1 - y)/(x + 4y - 3)$.
(d) Similarly, for $y' = (x + y)/(x - y + 1)$.
(e) Similarly, for $y' = (x - y - 4)/(x + y - 4)$.
(f) Devise a method of solution that will work in the exceptional case where $ae - bd = 0$, and apply it to the case $y' = (x + 2y - 1)/(2x + 4y - 1)$.

12. (*Algebraic, exponential, and explosive growth*) We saw, in Section 2.3.3, that the population model

$$\frac{dN}{dt} = \kappa N \qquad (\kappa > 0) \tag{12.1}$$

gives exponential growth, whereby $N \to \infty$ as $t \to \infty$. More generally, consider the model

$$\frac{dN}{dt} = \kappa N^p, \qquad (\kappa > 0) \tag{12.2}$$

where p is a positive constant. Solve (12.2) and show that if $0 < p < 1$ then the solution exhibits **algebraic growth** [i.e., $N(t) \sim \alpha t^\beta$ as $t \to \infty$]. Show that as $p \to 0$ the exponent β tends to unity, and as $p \to 1$ the exponent β tends to infinity. (Of course, when $p = 1$ we then have **exponential growth**, as mentioned above, so we can think – crudely – of exponential growth as a limiting case of algebraic growth, in the limit as the exponent β becomes infinite. Thus, exponential growth is powerful indeed.) If p is increased beyond 1 then we expect the growth to be even more spectacular. Show that if $p > 1$ then the solution exhibits **explosive growth**, explosive in the sense that $N \to \infty$ in *finite* time, as $t \to T$, where

$$T = \frac{1}{\kappa(p - 1)N_0^{p-1}}; \tag{12.3}$$

N_0 denotes the initial value $N(0)$. Observe that T diminishes as p increases.

13. (*Nondimensionalization*) In Example 7 we nondimensionalized according to $\bar{t} = at$ and $\overline{N} = N/N_0$. Instead, nondimensionalize (47) according to $\bar{t} = at$ and $\overline{N} = bN/a$, and thus derive the solution

$$\overline{N}(\bar{t}) = \frac{\beta}{\beta + (1 - \beta)e^{-\bar{t}}},$$

where $\beta \equiv bN_0/a$. Sketch the graph of $\overline{N}(\bar{t})$ versus $\bar{t}$, for several different values of β, labeling any key value(s).

14. The initial-value problem

$$mx'' + cx' + kx = F \sin \omega t; \qquad x(0) = x_0, \quad x'(0) = x_0' \tag{14.1}$$

corresponding to a damped mechanical oscillator driven by the force $F \sin \omega t$, contains seven parameters: $m, c, k, F, \omega, x_0, x_0'$. Nondimensionalize (14.1). How many parameters are present in the nondimensionalized system?

2.5 Exact Equations and Integrating Factors

Thus far we have developed solution techniques for first-order differential equations that are linear or separable. In addition, Bernoulli, Riccati, Clairaut, homogeneous, and almost-homogeneous equations were discussed in the exercises. In this section we consider one more important case, equations that are "exact," and ones that are not exact but can be made exact.

First, let us review some information, from the calculus, about partial derivatives. Specifically, recall that the symbol $\dfrac{\partial^2 f}{\partial x \partial y}$ is understood to mean $\dfrac{\partial}{\partial x}\left(\dfrac{\partial f}{\partial y}\right)$. If we use the standard subscript notation instead, then this quantity would be expressed as f_{yx}, that is, $(f_y)_x$. Does the order of differentiation matter? That is, is $f_{yx} = f_{xy}$? It is shown in the calculus that a sufficient condition for f_{xy} to equal f_{yx} is that f_x, f_y, f_{yx}, and f_{xy} all be continuous within the region in question. These conditions are met so typically in applications, that in textbooks on engineering and science f_{xy} and f_{yx} are generally treated as indistinguishable. Here, however, we will treat them as equal only if we explicitly assume the continuity of f_x, f_y, f_{yx}, and f_{xy}.

2.5.1. Exact differential equations. To motivate the idea of exact equations, consider the equation

$$\frac{dy}{dx} = \frac{\sin y}{2y - x\cos y} \tag{1}$$

or, rewritten in differential form,

$$\sin y\, dx + (x\cos y - 2y)dy = 0. \tag{2}$$

If we notice that the left-hand side is the differential of $F(x, y) = x\sin y - y^2$, then (2) is simply $dF = 0$, which can be integrated to give $F =$ constant; that is,

$$F(x, y) = x\sin y - y^2 = C, \tag{3}$$

where C is an arbitrary constant of integration. Equation (3) gives the general solution to (1), in implicit form.

Really, our use of the differential form (2) begs justification since we seem to have thereby treated dy/dx as a fraction of computable quantities dy and dx, whereas it is actually the limit of a difference quotient. Such justification is possible, but it may suffice to note that the use of differentials is a matter of convenience and is not essential to the method. For instance, observe that if we write

$$\sin y + (x\cos y - 2y)\frac{dy}{dx} = 0 \tag{4}$$

in place of (2), to avoid any questionable use of differentials, then the left-hand side of (4) is the x derivative (total, not partial) of $F(x, y) = x\sin y - y^2$:

$$\frac{d}{dx}F(x, y(x)) = \frac{d}{dx}\left(x\sin y - y^2\right) = \sin y + (x\cos y - 2y)\frac{dy}{dx},$$

so $dF/dx = 0$. Integrating the latter gives $F(x, y) = x \sin y - y^2 = C$, just as before.

Thus, let us continue, without concern about manipulating dy/dx as though it were a fraction. Seeking to generalize the method outlined above, we consider the differential equation

$$\frac{dy}{dx} = -\frac{M(x, y)}{N(x, y)}, \tag{5}$$

where the minus sign is included so that when we re-express (5) in the differential form

$$M(x, y)dx + N(x, y)dy = 0, \tag{6}$$

then both signs on the left will be positive. It is important to be aware that in equation (5) y is regarded as a function of x, as is clear from the presence of the derivative dy/dx. That is, there is a hierarchy whereby x is the independent variable and y is the dependent variable. But upon re-expressing (5) in the form (6) we change our viewpoint and now consider x and y as having the same status; now they are both independent variables.

We observe that integration of (6) is simple if $Mdx + Ndy$ happens to be the differential of some function $F(x, y)$, for if there does exist a function $F(x, y)$ such that

$$dF(x, y) = M(x, y)dx + N(x, y)dy, \tag{7}$$

then (6) is

$$dF(x, y) = 0, \tag{8}$$

which can be integrated to give the general solution

$$F(x, y) = C, \tag{9}$$

where C is an arbitrary constant.

Given $M(x, y)$ and $N(x, y)$, suppose that there does exist an $F(x, y)$ such that $Mdx + Ndy = dF$. Then we say that $Mdx + Ndy$ is an **exact differential**, and that (6) is an **exact differential equation**. That case is of great interest because its general solution is given immediately, in implicit form, by (9).

Two questions arise. How do we determine if such an F exists and, if it does, then how do we find it? The first is addressed by the following theorem.

THEOREM 2.5.1 *Test for Exactness*
Let $M(x, y)$, $N(x, y)$, $\partial M/\partial y$, and $\partial N/\partial x$ be continuous within a rectangle R in the x, y plane. Then $Mdx + Ndy$ is an exact differential, in R, if and only if

$$\boxed{\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}} \tag{10}$$

everywhere in R.

Partial Proof: Let us suppose that $M dx + N dy$ is exact, so that there is an F satisfying (7). Then it must be true, according to the chain rule of the calculus, that

$$M = \frac{\partial F}{\partial x} \tag{11a}$$

and

$$N = \frac{\partial F}{\partial y} \tag{11b}$$

Differentiating (11a) partially with respect to y, and (11b) partially with respect to x, gives

$$M_y = F_{xy}, \tag{12a}$$

and

$$N_x = F_{yx}. \tag{12b}$$

Since M, N, M_y, and N_x have been assumed continuous in R, it follows from (11) and (12) that F_x, F_y, F_{xy}, and F_{yx} are too, so $F_{xy} = F_{yx}$. Then it follows from (12) that $M_y = N_x$, which is equation (10). Because of the "if and only if" wording in the theorem, we also need to prove the reverse: that the truth of (10) implies the existence of F. That part of the proof can be carried out using results established in Section 16.12, and will not be given here. ■

Actually R need not be a rectangle; it merely needs to be "simply connected," that is, a region without holes. Simple connectedness will be defined and used extensively in Chapter 16 on Field Theory.

Assuming that the conditions of the theorem are met, so that we are assured that such an F exists, how do we *find* F? We can find it by integrating (11a) with respect to x, and (11b) with respect to y. Let us illustrate the method by reconsidering the example given above.

EXAMPLE 1. Consider equation (1) once again, or, in differential form,

$$\sin y \, dx + (x \cos y - 2y) dy = 0. \tag{13}$$

First, we identify $M = \sin y$, and $N = x \cos y - 2y$. Clearly, M, N, M_y, and N_x are continuous in the whole plane, so we turn to the exactness condition (10): $M_y = \cos y$, and $N_x = \cos y$, so (10) is satisfied, and it follows from Theorem 2.5.1 that there does exist an $F(x, y)$ such that the left-hand side of (13) is dF. Next, we find F from (11):

$$\frac{\partial F}{\partial x} = \sin y, \tag{14a}$$

$$\frac{\partial F}{\partial y} = x \cos y - 2y. \tag{14b}$$

Integrating (14a) partially, with respect to x, gives

$$F(x, y) = \int \sin y \, \partial x = x \sin y + A(y), \tag{15}$$

where the $\sin y$ integrand was treated as a constant in the integration since it was a "partial integration" on x, holding y fixed [just as y was held fixed in computing $\partial F/\partial x$ in (14a)]. The constant of integration A must therefore be allowed to depend upon y since y was held fixed and was therefore constant. If you are not convinced of this point, observe that taking a partial x-derivative of (15) does indeed recover (14a).

Observe that initially $F(x, y)$ was unknown. The integration of (14a) reduced the problem from an unknown function F of x and y to an unknown function A of y alone. $A(y)$, in turn, can now be determined from (14b). Specifically, we put the right-hand side of (15) into the left-hand side of (14b) and obtain

$$x \cos y + A'(y) = x \cos y - 2y, \tag{16}$$

where the prime denotes d/dy. Cancelling terms gives $A'(y) = -2y$, so

$$A(y) = -\int 2y\, dy = -y^2 + B, \tag{17}$$

where this integration was not a "partial integration," it was an ordinary integration on y since $A'(y)$ was an ordinary derivative of A. Combining (17) and (15) gives

$$F(x, y) = x \sin y - y^2 + B = \text{constant}. \tag{18}$$

Finally, absorbing B into the constant, and calling the result C, gives the general solution

$$x \sin y - y^2 = C \tag{19}$$

of (1), in implicit form.

COMMENT 1. Be aware that the partial integration notation $\int(\ )\partial x$ and $\int(\ )\partial y$ is not standard; we use it here because we find it reasonable, and helpful in reminding us that any y's in the integrand of $\int(\ )\partial x$ are to be treated as constants, and likewise any for any x's in $\int(\ )\partial y$.

COMMENT 2. From (13) all the way through (19), x and y have been regarded as independent variables. With (19) in hand, we can now return to our original viewpoint of y being a function of x. We can, if possible, solve (19) by algebra for $y(x)$ [in this case it is not because (19) is transcendental], plot the result, and so on. Even failing to solve (19) for $y(x)$, we can nevertheless verify that $x \sin y - y^2 = C$ satisfies (1) by differentiating with respect to x. That step gives $\sin y + x(\cos y)y' - 2yy' = 0$ or $y' = (\sin y)/(2y - x \cos y)$, which does agree with (1).

COMMENT 3. It would be natural to wonder how this method can *fail* to work. That is, whether or not $M_y = N_x$, why can't we always successfully integrate (11) to find F? The answer is to be found in (16). For suppose (16) were $2x \cos y + A'(y) = x \cos y - 2y$ instead. Then the $x \cos y$ terms would not cancel, as they did in (16), and we would have $A'(y) = -x \cos y - 2y$, which is impossible because it expresses a relationship between x and y, whereas x and y are regarded here as independent variables. Thus, the cancellation of the $x \cos y$ terms in (16) was crucial and was not an accident; it was a consequence of the fact that M and N satisfied the exactness condition (10).

COMMENT 4. Though we used (14a) first, then (14b), the order is immaterial and could have been reversed. ▮

2.5.2. Integrating factors. It may be discouraging to realize that for any given pair of functions M and N it is unlikely that the exactness condition (10) will be satisfied. However, there is power available to us that we have not yet tapped, for even if M and N fail to satisfy (10), so that the equation

$$M(x,y)dx + N(x,y)dy = 0 \tag{20}$$

is not exact, it *may* be possible to find a multiplicative factor $\sigma(x,y)$ so that

$$\sigma(x,y)M(x,y)dx + \sigma(x,y)N(x,y)dy = 0 \tag{21}$$

is exact. That is, we seek a function $\sigma(x,y)$ so that the revised exactness condition

$$\boxed{\frac{\partial}{\partial y}(\sigma M) = \frac{\partial}{\partial x}(\sigma N)} \tag{22}$$

is satisfied. Of course, we need $\sigma(x,y) \neq 0$ for (21) to be equivalent to (20).

If we can find a $\sigma(x,y)$ satisfying (22), then we call it an **integrating factor** of (20) because then (21) is equivalent to $dF = 0$, for some $F(x,y)$, and $dF = 0$ can be integrated immediately to give the solution of the original differential equation as $F(x,y) =$ constant.

How do we find such a σ? It is any (nonzero) solution of (22), that is, of

$$\sigma_y M + \sigma M_y = \sigma_x N + \sigma N_x. \tag{23}$$

Of course, (23) is a first-order partial differential equation on σ, so we have made dubious headway: to solve our original first-order ordinary differential equation on $y(x)$, we now need to solve the first-order partial differential equation (23) on $\sigma(x,y)$!

However, perhaps an integrating factor σ can be found that is a function of x alone: $\sigma(x)$. Then (23) reduces to the differential equation

$$\sigma M_y = \frac{d\sigma}{dx}N + \sigma N_x$$

or

$$\frac{d\sigma}{dx} = \sigma\left(\frac{M_y - N_x}{N}\right), \tag{24}$$

which is separable. This idea succeeds if and only if the $(M_y - N_x)/N$ ratio on the right-hand side of (24) is a function of x only, for if it did contain any y dependence then (24) would amount to the impossible situation of a function of x equalling a function of x and y, where x and y are independent variables. Thus, if

$$\boxed{\frac{M_y - N_x}{N} = \text{function of } x \text{ alone},} \tag{25}$$

then integration of (24) gives

$$\boxed{\sigma(x) = e^{\int \frac{M_y - N_x}{N}\, dx}.} \tag{26}$$

Actually, the general solution of (24) includes an arbitrary constant factor, but that factor is inconsequential and can be taken to be 1. Also, remember that we need σ to be nonzero and we are pleased to see, a posteriori, that the σ given in (26) cannot equal zero because it is an exponential function.

If $(M_y - N_x)/N$ is not a function of x alone, then an integrating factor $\sigma(x)$ does not exist, but we can try to find σ as a function of y alone: $\sigma(y)$. Then (23) reduces to

$$\frac{d\sigma}{dy}M + \sigma M_y = \sigma N_x$$

or

$$\frac{d\sigma}{dy} = -\sigma\left(\frac{M_y - N_x}{M}\right),$$

which, again, is separable. If

$$\boxed{\frac{M_y - N_x}{M} = \text{function of } y \text{ alone},} \tag{27}$$

then

$$\boxed{\sigma(y) = e^{-\int \frac{M_y - N_x}{M}\, dy}.} \tag{28}$$

EXAMPLE 2. Consider the equation (already expressed in differential form)

$$dx + \left(3x - e^{-2y}\right) dy = 0. \tag{29}$$

Then $M = 1$ and $N = 3x - e^{-2y}$, so (10) is not satisfied and (29) is not exact. Seeking an integrating factor that is a function of x alone, we find that

$$\frac{M_y - N_x}{N} = \frac{0 - 3}{3x - e^{-2y}} \neq \text{function of } x \text{ alone}, \tag{30}$$

and conclude that $\sigma(x)$ is not possible. Seeking instead an integrating factor that is a function of y alone,

$$\frac{M_y - N_x}{M} = \frac{0 - 3}{1} = -3 = \text{function of } y \text{ alone}, \tag{31}$$

so that $\sigma(y)$ is possible, and is given by

$$\sigma(y) = e^{-\int \frac{M_y - N_x}{M}\, dy} = e^{\int 3\, dy} = e^{3y}. \tag{32}$$

Multiply (29) through by the integrating factor $\sigma = e^{3y}$ and obtain

$$e^{3y}dx + e^{3y}\left(3x - e^{-2y}\right) dy = 0,$$

which is now exact. Thus,

$$\frac{\partial F}{\partial x} = e^{3y}$$

and

$$\frac{\partial F}{\partial y} = e^{3y}\left(3x - e^{-2y}\right)$$

so

$$F(x,y) = \int e^{3y}\,\partial x = xe^{3y} + A(y),$$

and

$$\frac{\partial F}{\partial y} = e^{3y}\left(3x - e^{-2y}\right) = 3xe^{3y} + A'(y). \tag{33}$$

The latter gives

$$A'(y) = -e^{y}$$

so

$$A(y) = -e^{y} + B.$$

Thus,

$$F(x,y) = xe^{3y} + A(y) = xe^{3y} + (-e^{y} + B) = \text{constant}$$

or

$$xe^{3y} - e^{y} = C, \tag{34}$$

where C is an arbitrary constant; (34) is the general solution of (29), in implicit form.

COMMENT. Can we solve (34) for y? If we let $e^{y} \equiv z$, then (34) is the cubic equation $xz^{3} - z = C$ in z, and there is a known solution to cubic equations. If we can solve for z, then we have y as $y = \ln z$. However, the solution of that cubic equation (as can be obtained using the *Maple* solve command) is quite a messy expression. ▮

EXAMPLE 3. *First-Order Linear Equation.* We've already solved the general first-order linear equation

$$\frac{dy}{dx} + p(x)y = q(x) \tag{35}$$

in Section 2.2, but let us see if we can solve it again, using the ideas of this section. First, express (35) in the form

$$[p(x)y - q(x)]\,dx + dy = 0. \tag{36}$$

Thus, $M = p(x)y - q(x)$ and $N = 1$, so $M_y = p(x)$ and $N_x = 0$. Hence $M_y \neq N_x$, so (36) is not exact [except in the trivial case when $p(x) = 0$]. Since

$$\frac{M_y - N_x}{N} = \frac{p(x) - 0}{1} = \text{function of } x \text{ alone},$$

$$\frac{M_y - N_x}{M} = \frac{p(x) - 0}{p(x)y - q(x)} \neq \text{function of } y \text{ alone},$$

we can find an integrating factor that is a function of x alone, but not one that is a function of y alone. We leave it for the exercises to show that the integrating factor is

$$\sigma(x) = e^{\int p(x)\,dx},$$

and that the final solution (this time obtainable in explicit form) is

$$y(x) = e^{-\int p\,dx}\left(\int e^{\int p\,dx} q\,dx + C\right), \tag{37}$$

as found earlier, in Section 2.2. ∎

Closure. Let us summarize the main results. Given a differential equation $dy/dx = f(x, y)$, the first step in using the method of exact differentials is to re-express it in the differential form $M(x, y)dx + N(x, y)dy = 0$. If M, N, M_y, and N_x are all continuous in the region of interest, check to see if the exactness condition (10) is satisfied. If it is, then the equation is exact, and its general solution is $F(x, y) = C$, where F is found by integrating (11a) and (11b). As a check on your work, a differential of $F(x, y) = C$ should give you back the original equation $Mdx + Ndy = 0$.

If it is not exact, see if $(M_y - N_x)/N$ is a function of x alone. If it is, then an integrating factor $\sigma(x)$ can be found from (26). Multiplying the given equation $Mdx + Ndy = 0$ through by that $\sigma(x)$, the new equation is exact, and you can proceed as outlined above for an exact equation.

If $(M_y - N_x)/N$ is not a function of x alone, check to see if $(M_y - N_x)/M$ is a function of y alone. If it is, then an integrating factor $\sigma(y)$ can be found from (28). Multiplying $Mdx + Ndy = 0$ through by that $\sigma(y)$, the new equation is exact, and you can proceed as outlined above for an exact equation.

If $M_y \neq N_x$, $(M_y - N_x)/N$ is not a function of x alone, and $(M_y - N_x)/M$ is not a function of y alone, then the method is of no help unless an integrating factor σ can be found that is a function of both x and y.

EXERCISES 2.5

NOTE: Solutions should be expressed in explicit form if possible.

1. Show that the equation is exact, and obtain its general solution. Also, find the particular solution corresponding to the given initial condition as well.

(a) $3dx - dy = 0$; $y(0) = 6$
(b) $x^2dx + y^2dy = 0$; $y(9) = -1$
(c) $x\,dx + 2ydy = 0$; $y(1) = 2$
(d) $4\cos 2u\,du - e^{-5v}dv = 0$; $v(0) = -6$
(e) $e^ydx + (xe^y - 1)dy = 0$; $y(-5) = 6$
(f) $(e^y + z)dy - (\sin z - y)dz = 0$; $z(0) = 0$
(g) $(x - 2z)dx - (2x - z)dz = 0$; $z(-3) = 5$
(h) $(\sin y + y\cos x)dx + (\sin x + x\cos y)dy = 0$; $y(2) = 3$
(i) $(\sin xy + xy\cos xy)dx + x^2\cos xy\,dy = 0$; $y(0) = -1$
(j) $(3x^2\sin 2y - 2xy)dx + (2x^3\cos 2y - x^2)dy = 0$; $y(0.5) = 3.1$
(k) $(4x^3y^5\sin 3x + 3x^4y^5\cos 3x)dx + 5x^4y^4\sin 3x\,dy = 0$; $y(0) = 1$
(l) $3x^2y\ln y\,dx + (x^3\ln y + x^3 - 2y)dy = 0$; $y(8) = -3$
(m) $(2ye^{2xy}\sin x + e^{2xy}\cos x + 1)dx + 2xe^{2xy}\sin x\,dy = 0$; $y(2.3) = -1.25$

2.(a)–(m) Find the general solution of the equation given in Exercise 1 using computer software, and also the particular solution corresponding to the given initial condition.

3. Make up three different examples of exact equations.

4. Determine whatever conditions, if any, are needed on the constants $a, b, \ldots, f, A, B, \ldots, F$ for the equation to be exact.

(a) $(ax + by + c)dx + (Ax + By + C)dy = 0$
(b) $(ax^2 + by^2 + cxy + dx + ey + f)dx + (Ax^2 + By^2 + Cxy + Dx + Ey + F)dy = 0$

5. Find a suitable integrating factor $\sigma(x)$ or $\sigma(y)$, and use it to find the general solution of the differential equation.

(a) $3y dx + dy = 0$
(b) $y dx + x \ln x \, dy = 0$
(c) $y \ln y \, dx + (x + y) dy = 0$
(d) $dx + (x - e^{-y}) dy = 0$
(e) $dx + x dy = 0$
(f) $(ye^{-x} + 1) dx + (xe^{-x}) dy = 0$
(g) $\cos y \, dx - [2(x - y) \sin y + \cos y] dy = 0$
(h) $(1 - x - z) dx + dz = 0$
(i) $(2 + \tan^2 x)(1 + e^{-y}) dx - e^{-y} \tan x \, dy = 0$
(j) $(3u^2 \sinh 3v - 2u) du + 3u^3 \cosh 3v \, dv = 0$
(k) $\cos x \, dx + (3 \sin x + 3 \cos y - \sin y) dy = 0$
(l) $(y \ln y + 2xy^2) dx + (x + x^2 y) dy = 0$
(m) $(3x - 2p) dx - x \, dp = 0$
(n) $y \, dx + (x^2 - x) dy = 0$
(o) $2xy \, dx + (y^2 - x^2) dy = 0$

6. (*First-order linear equation*) Verify that $\sigma(x) = e^{\int p(x) \, dx}$ is an integrating factor for the general linear first-order equation (35), and use it to derive the general solution (37).

7. Show that the given equation is not exact and that an integrating factor depending on x alone or y alone does not exist. If possible, find an integrating factor in the form $\sigma(x, y) = x^a y^b$, where a and b are suitably chosen constants. If such a σ can be found, then use it to obtain the general solution of the differential equation; if not, state that.

(a) $(3xy - 2y^2) dx + (2x^2 - 3xy) dy = 0$
(b) $(3xy + 2y^2) dx + (3x^2 + 4xy) dy = 0$
(c) $(x + y^2) dx + (x - y) dy = 0$
(d) $y dx - (x^2 y - x) dy = 0$

8. Show that the equation is not exact and that an integrating factor depending on x alone or y alone does not exist. Nevertheless, find a suitable integrating factor by inspection, and use it to obtain the general solution.

(a) $e^y dx + e^x dy = 0$ (b) $y^2 dx - e^{3x} dy = 0$

(c) $e^{2y} dx - \tan x \, dy = 0$

9. Obtain the general solution, using the methods of this section.

(a) $\dfrac{dy}{dx} = \dfrac{x - y}{x + y}$ (b) $\dfrac{dr}{d\theta} = -\dfrac{r^2 \cos \theta}{2r \sin \theta + 1}$

(c) $\dfrac{dy}{dx} = \dfrac{2xy - e^y}{x(e^y - x)}$ (d) $\dfrac{dy}{dx} = \dfrac{y(2x - \ln y)}{x}$

(e) $\dfrac{dy}{dx} = -\dfrac{\sin y + y \cos x}{\sin x + x \cos y}$

10. What do the integrating factors defined by (26) and (28) turn out to be if the equation is exact to begin with?

11.(a) Show that $(x^3 + y) dx + (y^3 + x) dy = 0$ is exact.
(b) More generally, is $M(x, y) dx + M(y, x) dy$ exact? Explain.

12. If $F(x, y) = C$ is the general solution (in implicit form) of a given first-order equation, then what is the particular solution (in implicit form) satisfying the initial condition $y(a) = b$?

13. If $M dx + N dy = 0$ and $P dx + Q dy = 0$ are exact, is $(M + P) dx + (N + Q) dy = 0$ exact? Explain.

14. Show that for $[p(x) + q(y)] dx + [r(x) + s(y)] dy = 0$ to be exact, it is necessary and sufficient that $q(y) dx + r(x) dy$ be an exact differential.

15. Show that for $p(x) dx + q(x) r(y) dy = 0$ to be exact, it is necessary and sufficient that $q(x)$ be a constant.

Chapter 2 Review

Following is a listing of the types of equations covered in this chapter.

SECTION 2.2

First-order linear: $y' + p(x)y = q(x)$.

This equation can be solved by the integrating factor method or by solving the homogeneous equation and using variation of parameters. Its general solution is

$$y(x) = e^{-\int p(x)\,dx}\left(\int e^{\int p(x)\,dx} q(x)\,dx + C\right).$$

A particular solution satisfying $y(a) = b$ is

$$y(x) = e^{-\int_a^x p(\xi)\,d\xi}\left(\int_a^x e^{\int_a^\xi p(\zeta)\,d\zeta} q(\xi)\,d\xi + b\right).$$

Bernoulli: $y' + p(x)y = q(x)y^n$. $(n \neq 0, 1)$.

This equation can be converted to the first-order *linear* equation $v' + (1-n)p(x)v = (1-n)q(x)$ by the change of variables $v = y^{1-n}$ (Exercise 9).

Riccati: $y' = p(x)y^2 + q(x)y + r(x)$.

This equation can be solved by setting $y = Y(x) + \dfrac{1}{v}$, if a particular solution $Y(x)$ of the Riccati equation can be found (Exercise 11).

d'Alembert–Lagrange: $y' = xf(y') + g(y')$. $[f(y') \neq y']$

By letting $y' = p$ be a new independent variable, one can obtain a *linear* first-order equation on $x(p)$ (Exercise 13).

Clairaut: $y' = xy' + g(y')$.

This equation admits the family of straight-line solutions $y = Cx + g(C)$ and, in general, a singular solution as well (Exercise 14).

SECTION 2.4

Separable: $y' = X(x)Y(y)$.

General solution obtained by integrating

$$\int \frac{dy}{Y(y)} = \int X(x)dx.$$

Homogeneous: $y' = f\left(\frac{y}{x}\right)$.

Can be made separable by setting $v = y/x$ (Exercise 9).

Almost Homogeneous: $y' = \frac{ax + by + c}{dx + ey + f}$. $(ae - bd \neq 0)$

Can be made homogeneous by setting $x = u + h$, $y = v + k$ (Exercise 11).

SECTION 2.5

Exact: $M(x, y)dx + N(x, y)dy = 0$. $(M_y = N_x)$

General solution $F(x, y) = C$ found by integrating $F_x = M$, $F_y = N$. If $M_y \neq N_x$, can make exact by means of an integrating factor $\sigma(x)$ if $(M_y - N_x)/N$ is a function of x only, or by an integrating factor $\sigma(y)$ if $(M_y - N_x)/M$ is a function of y only.

Chapter 3

Linear Differential Equations of Second Order and Higher

PREREQUISITES: In this chapter on linear differential equations, we encounter systems of linear algebraic equations, and it is presumed that the reader is familiar with the theory of the existence and uniqueness of solutions to such equations, especially as regards the role of the determinant of the coefficient matrix. That material is covered in Chapters 8–10, but the essential results that are needed for the present chapter are summarized briefly in Appendix B. Thus, either Sections 8.1–10.6 or Appendix B is a prerequisite for this chapter. Also presumed is a familiarity with the complex plane and the algebra of complex numbers. That material is covered in Section 21.2 which, likewise, is a prerequisite for Chapter 3.

3.1 Introduction

As we prepare to move from first-order equations to those of higher order, this is a good time to pause for an overview that looks back to Chapter 2 and ahead to Chapters 3–7. If, as you proceed through Chapters 3–7, you lose sight of the forest for the trees, we urge you to come back to this overview.

LINEAR EQUATIONS

First order:

$$y' + p(x)y = q(x). \tag{1}$$

General solution found [(2.1) in Section 2.2] in explicit form. Existence and uniqueness of solution of initial-value problem [with $y(a) = b$] guaranteed over a predetermined interval, based upon the continuity of $p(x)$ and $q(x)$. Solution of initial-value problem expressible as a superposition of responses to the two inputs [the initial value b and the forcing function $q(x)$] with each response being proportional to that input: for example, if we double the input we double the output.

Second order and higher:

$$a_0(x)\frac{d^ny}{dx^n} + a_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{dy}{dx} + a_n(x)y = f(x). \qquad (2)$$

Constant coefficients (the a_j's are constants) and homogeneous ($f = 0$):
This is the simplest case. We will see (Section 3.4) that the general solution can be found in terms of exponential functions, and perhaps powers of x times exponential functions.

Constant coefficients and nonhomogeneous:
Additional solution is needed due to the forcing function $f(x)$ and can be found by the method of undetermined coefficients (Section 3.7.2) or the method of variation of parameters (Sections 3.7.3 and 3.7.4). Still simple. An alternative approach, the Laplace transform, is given in Chapter 5.

Nonconstant coefficients:
Essentially, the only simple case is the Cauchy – Euler equation (Section 3.6.1). Other cases are so much more difficult that we give up on finding closed form solutions and use power series methods (Chapter 4). Two particularly important cases are the Legendre (Section 4.4) and Bessel (Section 4.6) equations, which will be needed later in the chapters on partial differential equations.

NONLINEAR EQUATIONS

First order:

$$y' = f(x, y). \qquad (3)$$

No solution available for the general case. Need to identify subcategories that are susceptible to special solution techniques. The most important of these subcategories are separable equations (Section 2.4) and exact equations (Section 2.5), and these methods give solutions in implicit form. Several important but more specialized cases are given in the exercises: the Bernoulli, Riccati, d'Alembert – Lagrange, and Clairaut equations in Section 2.2, and "homogeneous" equations in Section 2.4. The idea of the response being a superposition of responses, as it is for the linear equation, is not applicable for nonlinear equations. The subcategories and special cases mentioned above by no means cover all possible equations of the form $y' = f(x, y)$, so that many first-order nonlinear equations simply are not solvable by any known means. A powerful alternative to analytical methods, [i.e., methods

> designed to obtain an analytical expression for $y(x)$], is to seek a solution in numerical form, with the help of a computational algorithm and a computer, and these methods are discussed in Chapter 6.

Second order and higher:

> Some nonlinear equations of first order can be solved analytically, as we have seen, but for nonlinear equations of higher order analytical solution is generally out of the question, and we rely instead upon a blend of numerical solution (Chapter 6) and qualitative methods, such as the phase plane method described in Chapter 7.

To get started, we limit our attention in the next several sections to the **homogeneous** version of the linear equation (2), namely, where $f(x) = 0$, because that case is simpler and because to solve the nonhomogeneous case we will need to solve the homogeneous version first, anyhow.

To attach physical significance to the distinction between homogeneous and nonhomogeneous equations, it may help to recall from Section 1.3 that the differential equation governing a mechanical oscillator is

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t), \tag{4}$$

where m, c, k are the mass, damping coefficient, and spring stiffness, respectively, and $F(t)$ is the applied force. (In this case, of course, the variables happen to be x and t rather than y and x.) If $F(t) = 0$, then (4) governs the unforced, or "free," vibration of the mass m. Likewise, for any linear differential equation, if all terms containing the unknown and its derivatives are moved to the left-hand side, then whatever is left on the right-hand side is regarded as a "forcing function." From a physical point of view then, when we consider the homogeneous case in the next several sections, we are really limiting our attention to unforced systems.

A brief outline of this chapter follows:

3.2 *Linear Dependence and Linear Independence.* The concept of a general solution to a linear differential equation requires the idea of linear dependence and linear independence, so these ideas are introduced first.

3.3 *Homogeneous Equation: General Solution.* Here we establish the concept of a general solution to the homogeneous equation (2), but do not yet show how to obtain it.

3.4 *Solution of Homogeneous Equation: Constant Coefficients.* It is shown how to find the general solution in the form of a linear combination of solutions that are either exponentials or powers of x times exponentials.

3.5 *Application to Harmonic Oscillator: Free Oscillation.* The foregoing concepts and methods are applied to an extremely important physical application: the free oscillation of a harmonic oscillator.

3.6 *Solution of Homogeneous Equation: Nonconstant Coefficients.* Nonconstant-coefficient equations can be solved in closed form only in exceptional cases. The most important such case is the Cauchy–Euler equation, and that case occupies most of this section.

3.7 *Solution of Nonhomogeneous Equation.* It is shown how to find the additional solution, due to the forcing function, by the methods of undetermined coefficients and variation of parameters.

3.8 *Application to Harmonic Oscillator: Forced Oscillation.* We return to the example of the harmonic oscillator, begun in Section 3.5, and obtain and discuss the solution for the forced oscillation.

3.9 *Systems of Linear Differential Equations.* We consider linear systems of n coupled first-order differential equations on n unkowns and show how to obtain uncoupled nth-order differential equations on each of the n unknowns, which equations can then be solved by the methods described in the preceding sections of this chapter.

3.2 Linear Dependence and Linear Independence

Asked how many different paints he had, a painter replied five: red, blue, green, yellow, and purple. However, it could be argued that the count was inflated since only three (for instance red, blue, and yellow) are independent: the green can be obtained from a certain proportion of the blue and the yellow, and the purple can be obtained from the red and the blue. Similarly, in studying linear differential equations, we will need to determine how many "different," or "independent," functions are contained within a given set of functions. The concept is made precise as follows. We begin by defining a **linear combination** of a set of functions $f_1, \dots, f_k$ as any function of the form $\alpha_1 f_1 + \cdots + \alpha_k f_k$, where the α_j's are constants. For instance, $2 \sin x - \pi \cos x$ is a linear combination of $\sin x$ and $\cos x$.

DEFINITION 3.2.1 *Linear Dependence and Linear Independence*
A set of functions $\{u_1, \dots, u_n\}$ is said to be **linearly dependent** on an interval I if at least one of them can be expressed as a linear combination of the others on I. If none can be so expressed, then the set is **linearly independent**.

If we do not specify the interval I, then it will be understood to be the entire x axis. NOTE: Since the terms linearly dependent and linearly independent will appear repeatedly, it will be convenient to abbreviate them in this book as **LD** and

LI, respectively, but be aware that this notation is not standard outside of this text.

EXAMPLE 1. The set $\{x^2, e^x, e^{-x}, \sinh x\}$ is seen to be LD (linearly dependent) because we can express $\sinh x$ as a linear combination of the others:

$$\sinh x = \frac{e^x - e^{-x}}{2} = \frac{1}{2}e^x - \frac{1}{2}e^{-x} + 0\,x^2. \tag{1}$$

In fact, we could express e^x as a linear combination of the others too, for solving (1) for e^x gives $e^x = 2\sinh x + e^{-x} + 0\,x^2$. Likewise, we could express $e^{-x} = e^x - 2\sinh x + 0\,x^2$. We cannot express x^2 as a linear combination of the others [since we cannot solve (1) for x^2], but the set is LD nonetheless, because we only need to be able to express "at least one" member as a linear combination of the others. NOTE: The hyperbolic sine and cosine functions, $\sinh x$ and $\cosh x$, were studied in the calculus, but if these functions and their graphs and properties are not familiar to you, you may wish to turn to the review in Section 3.4.1. ▮

The foregoing example was simple enough to be worked by inspection. In more complicated cases, the following theorem provides a test for determining whether a given set is LD or LI.

THEOREM 3.2.1 *Test for Linear Dependence/Independence*
A finite set of functions $\{u_1, \ldots, u_n\}$ is LD on an interval I if and only if there exist scalars α_j, not all zero, such that

$$\alpha_1 u_1(x) + \alpha_2 u_2(x) + \cdots + \alpha_n u_n(x) = 0 \tag{2}$$

identically on I. If (2) is true only if all the α's are zero, then the set is LI on I.

Proof: Because of the "if and only if" we need to prove the statement in both directions. First, suppose that the set is LD. Then, according to the definition of linear dependence, one of the functions, say u_j, can be expressed as a linear combination of the others:

$$u_j(x) = \alpha_1 u_1(x) + \cdots + \alpha_{j-1}u_{j-1}(x) + \alpha_{j+1}u_{j+1}(x) + \cdots + \alpha_n u_n(x), \tag{3}$$

which equation can be rewritten as

$$\alpha_1 u_1(x) + \cdots + \alpha_{j-1}u_{j-1}(x) + (-1)u_j(x) + \alpha_{j+1}u_{j+1}(x) + \cdots + \alpha_n u_n(x) = 0. \tag{4}$$

Even if all the other α's are zero, the coefficient α_j of $u_j(x)$ in (4) is nonzero, namely, -1, so there do exist scalars $\alpha_1, \ldots, \alpha_n$ not all zero such that (2) holds.

Conversely, suppose that (2) holds with the α's not all zero. If α_k, for instance, is nonzero, then (2) can be divided by α_k and solved for $u_k(x)$ as a linear combination of the other u's, in which case $\{u_1, \dots, u_n\}$ is LD. ■

EXAMPLE 2. To determine if the set $\{1, x, x^2\}$ is LD or LI using Theorem 3.2.1, write equation (2),

$$\alpha_1 + \alpha_2 x + \alpha_3 x^2 = 0, \tag{5}$$

and see if the truth of (5) requires all the α's to be zero. Since (5) needs to hold for all x's in the interval (which we take to be $-\infty < x < \infty$), let us write it for $x = 0, 1, 2$, say, to generate three equations on the three α's:

$$\begin{aligned} \alpha_1 &= 0, \\ \alpha_1 + \alpha_2 + \alpha_3 &= 0, \\ \alpha_1 + 2\alpha_2 + 4\alpha_3 &= 0. \end{aligned} \tag{6}$$

Solution of (6) gives $\alpha_1 = \alpha_2 = \alpha_3 = 0$, so the set is LI.

In fact, (5) really amounts to an *infinite* number of linear algebraic equations on the three α's since there is no limit to the number of x values that could be chosen. However, three different x values sufficed to establish that all of the α's must be zero. ■

Alternative to writing out (2) for n specific x values, to generate n equations on $\alpha_1, \dots, \alpha_n$, it is more common to generate n such equations by writing (2) and its first $n - 1$ derivatives (assuming, of course, that $u_1, \dots, u_n$ are $n - 1$ times differentiable on I),

$$\begin{aligned} \alpha_1 u_1(x) + \cdots + \alpha_n u_n(x) &= 0, \\ \alpha_1 u_1'(x) + \cdots + \alpha_n u_n'(x) &= 0, \\ &\vdots \\ \alpha_1 u_1^{(n-1)}(x) + \cdots + \alpha_n u_n^{(n-1)}(x) &= 0. \end{aligned} \tag{7}$$

Let us denote the determinant of the coefficients as

$$\boxed{W\left[u_1, \dots, u_n\right](x) = \begin{vmatrix} u_1(x) & \cdots & u_n(x) \\ u_1'(x) & \cdots & u_n'(x) \\ \vdots & & \vdots \\ u_1^{(n-1)}(x) & \cdots & u_n^{(n-1)}(x) \end{vmatrix},} \tag{8}$$

which is known as the **Wronskian determinant** of $u_1, \dots, u_n$, or simply the **Wronskian** of $u_1, \dots, u_n$, after the Polish mathematician *Josef M. H. Wronski* (1778–1853). The Wronskian W is itself a function of x.

From the theory of linear algebraic equations, we know that if there is any value of x in I, say x_0, such that $W\left[u_1, \dots, u_n\right](x_0) \neq 0$, then it follows from (7) with x set equal to x_0, that all the α's must be zero, so the set $\{u_1, \dots, u_n\}$ is LI.

THEOREM 3.2.2 *Wronskian Condition for Linear Independence*
If, for a set of functions $\{u_1, \ldots, u_n\}$ having derivatives through order $n-1$ on an interval I, $W[u_1, \ldots, u_n](x)$ is not identically zero on I, then the set is LI on I.

Be careful not to read into Theorem 3.2.2 a converse, namely, that if $W[u_1, \ldots, u_n](x)$ *is* identically zero on I (which we write as $W \equiv 0$), then the set is LD on I. In fact, the latter is not true, as shown by the following example.

EXAMPLE 3. Consider the set $\{u_1, u_2\}$, where

$$u_1(x) = \begin{cases} x^2, & x \le 0 \\ 0, & x \ge 0, \end{cases} \qquad u_2(x) = \begin{cases} 0, & x \le 0 \\ x^2, & x \ge 0. \end{cases} \tag{9}$$

(Sketch their graphs.) Then (2) becomes

$$\begin{aligned} &\alpha_1 x^2 + \alpha_2(0) = 0 \qquad \text{for } x \le 0 \\ &\alpha_1(0) + \alpha_2 x^2 = 0 \qquad \text{for } x \ge 0. \end{aligned}$$

The first implies that $\alpha_1 = 0$, and the second implies that $\alpha_2 = 0$. Hence $\{u_1, u_2\}$ is LI. Yet, $W[u_1, u_2](x) = \begin{vmatrix} x^2 & 0 \\ 2x & 0 \end{vmatrix} = 0$ on $x \le 0$, and $W[u_1, u_2](x) = \begin{vmatrix} 0 & x^2 \\ 0 & 2x \end{vmatrix} = 0$ on $x \ge 0$, so $W[u_1, u_2](x) \equiv 0$ for all x. ▮

However, our interest in linear dependence and independence, in this chapter, is not going to be in connection with sets of randomly chosen functions, but with sets of functions which have in common that they are solutions of a given linear homogeneous differential equation. In that case, it can be shown that the inverse of Theorem 3.2.2 *is* true: that is, if $W \equiv 0$, then the set is LD. Thus, for that case we have the following stronger theorem which, for our subsequent purposes, will be more important to us than Theorem 3.2.2.

THEOREM 3.2.3 *A Necessary and Sufficient Condition for Linear Dependence*
If $u_1, \ldots, u_n$ are solutions of an nth-order linear homogeneous differential equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{10}$$

where the coefficients $p_j(x)$ are continuous on an interval I, then $W[u_1, \ldots, u_n](x) \equiv 0$ on I is both necessary and sufficient for the linear dependence of the set $\{u_1, \ldots, u_n\}$ on I.

EXAMPLE 4. It is readily verified that each of the functions $1, e^x, e^{-x}$ satisfies the equation $y''' - y' = 0$. Since their Wronskian is

$$W\left[1, e^x, e^{-x}\right](x) = \begin{vmatrix} 1 & e^x & e^{-x} \\ 0 & e^x & -e^{-x} \\ 0 & e^x & e^{-x} \end{vmatrix} = 2 \neq 0,$$

it follows from Theorem 3.2.3 that the set $\{1, e^x, e^{-x}\}$ is LI. Another set of solutions of $y''' - y' = 0$ is $e^x, e^{-x}, \cosh x$. Their Wronskian is

$$W\left[e^x, e^{-x}, \cosh x\right](x) = \begin{vmatrix} e^x & e^{-x} & \cosh x \\ e^x & -e^{-x} & \sinh x \\ e^x & e^{-x} & \cosh x \end{vmatrix} = 0,$$

so the set $\{e^x, e^{-x}, \cosh x\}$ is LD. ▮

In connection with Theorem 3.2.3, it would be natural to wonder if W could be zero for some x's and nonzero for others. Subject to the conditions of that theorem, it can be shown (Exercise 5) that

$$W(x) = W(\xi) \exp\left[-\int_\xi^x p_1(t)\, dt\right], \tag{11}$$

where ξ is any point in the interval and p_1 is the coefficient of the next-to-highest derivative in (10), and where we have written $W[u_1, \ldots, u_n](x)$ as $W(x)$, and $W[u_1, \ldots, u_n](\xi)$ as $W(\xi)$, for brevity. Due to the French mathematican *Joseph Liouville* (1809–1882), and known as **Liouville's formula**, (11) shows that under the conditions of Theorem 3.2.3 the Wronskian is either everywhere zero or everywhere nonzero, for the exponential function is positive for all finite values of its argument and the constant $W(\xi)$ is either zero or not. This fact is illustrated by the two Wronskians in Example 4.

Finally, it is useful to cite the following three simple results, proofs of which are left for the exercises.

THEOREM 3.2.4 *Linear Dependence/Independence of Two Functions*
A set of two functions, $\{u_1, u_2\}$, is LD if and only if one is expressible as a scalar multiple of the other.

THEOREM 3.2.5 *Linear Dependence of Sets Containing the Zero Function*
If a set $\{u_1, \ldots, u_n\}$ contains the zero function [that is, $u_j(x) = 0$ for some j], then the set is LD.

THEOREM 3.2.6 *Equating Coefficients*
Let $\{u_1, \ldots, u_n\}$ be LI on an interval I. Then, for

$$a_1u_1(x) + \cdots + a_nu_n(x) = b_1u_1(x) + \cdots + b_nu_n(x)$$

to hold on I, it is necessary and sufficient that $a_j = b_j$ for each $j = 1, \ldots, n$. That is, the coefficients of corresponding terms on the left- and right-hand sides must match.

EXAMPLE 5. The set $\{x, \sin x\}$ is LI on $-\infty < x < \infty$ according to Theorem 3.2.4 because x is surely not expressible as a constant times $\sin x$ (for $x/\sin x$ is not a constant), nor is $\sin x$ expressible as a constant times x. ▮

EXAMPLE 6. We've seen that $\{1, e^x, e^{-x}\}$ is LI on $-\infty < x < \infty$. Thus, if we meet the equation

$$a + be^x + ce^{-x} = 6 - 2e^{-x}, \tag{12}$$

then it follows from Theorem 3.2.6 that we must have $a = 6$, $b = 0$, $c = -2$, for if we rewrite (12) as

$$(a - b)(1) + be^x + (c + 2)e^{-x} = 0,$$

then it follows from the linear independence of $1, e^x, e^{-x}$ that $a - 6 = 0$, $b = 0$, $c + 2 = 0$; that is, $a = 6$, $b = 0$, $c = -2$. ▮

Closure. We have introduced the concept of linear dependence and linear independence as preliminary to our development of the theory of linear differential equations, which follows next. Following the definitions of these terms, we gave three theorems for the testing of a given set of functions to determine if they are LI or LD. Of these, Theorem 3.2.3 will be most useful to us in the sections to follow because it applies to sets of functions that arise as solutions of a given differential equation.

In case you have trouble remembering which of the conditions $W = 0$ and $W \neq 0$ corresponds to linear dependence and which to linear independence, think of it this way. If we randomly make up a determinant, the chances are that its value is nonzero; that is the generic case. Likewise, if we randomly select a set of functions out of the set of all possible functions, the generic case is for them to be unrelated – namely, LI. The generic cases go together ($W \neq 0$ corresponding to linear independence) and the nongeneric cases go together ($W = 0$ corresponding to linear dependence).

The concept of linear dependence and independence will prove to be important to us later as well, when we study n-dimensional vector spaces and the expansion of a given vector in terms of a set of base vectors.

EXERCISES 3.2

1. (a) Can a set be neither LD nor LI? Explain.
(b) Can a set be both LD and LI? Explain.

2. Show that the following sets are LD by expressing one of the functions as a linear combination of the others.

(a) $\{1,\ x+2,\ 3x-5\}$
(b) $\{x^2,\ x^2+x,\ x^2+x+1,\ x-1\}$
(c) $\{x^4+x^2+1,\ x^4-x^2+1,\ x^4-x^2-1\}$
(d) $\{e^x,\ e^{2x},\ \sinh x,\ \cosh x\}$
(e) $\{\sinh 3x,\ e^x,\ e^{3x},\ e^{5x},\ e^{-3x}\}$
(f) $\{e^x,\ e^{2x},\ xe^x,\ (7x-2)e^x\}$
(g) $\{0, x, x^3\}$
(h) $\{x, 2x, x^2\}$

3. Show whether the given set is LD or LI. HINT: In most cases, the brief discussion of determinants given in Appendix B will suffice; in others, you will need to use known properties of determinants given in Section 10.4. Also, note that the *Maple* command for calculating determinants (the elements of which need not be constants) is given at the end of Section 10.4.

(a) $\{1,\ x,\ x^2, \ldots, x^n\}$
(b) $\{e^{a_1x},\ e^{a_2x}, \ldots, e^{a_nx}\}$
(c) $\{1,\ 1+x,\ 1+x^2\}$
(d) $\{e^x,\ e^{2x}\}$
(e) $\{\sin x,\ \cos x,\ \sinh x\}$
(f) $\{x,\ x^2\}$
(g) $\{1,\ \sin 3x\}$
(h) $\{1-x,\ 1+x,\ x^3\}$
(i) $\{x,\ 1+x,\ e^x\}$
(j) $\{x,\ e^x,\ \cos x\}$

4. Verify that each of the given functions is a solution of the given differential equation, and then use Theorem 3.2.3 to determine if the set is LD or LI. As a check, use Theorem 3.2.4 if that theorem applies.

(a) $y''' - 6y'' + 11y' - 6y = 0, \quad \{e^x, e^{2x}, e^{3x}\}$
(b) $y'' + 4y = 0, \quad \{\sin 2x, \cos 2x\}$
(c) $y''' - 6y'' + 9y' - 4y = 0, \quad \{e^x, xe^x, e^{4x}\}$
(d) $y''' - 6y'' + 9y' - 4y = 0, \quad \{e^x, xe^x, (1-x)e^x\}$
(e) $y''' - y'' - 2y' = 0, \quad \{1, e^{-x}, e^{2x}\}$
(f) $y'''' - 5y'' + 4y = 0, \quad \{e^x, e^{-x}, e^{2x}, e^{-2x}\}$
(g) $x^2y'' - 3xy' + 3y = 0, \quad \{x, x^3\}, \quad$ on $x > 0$
(h) $x^2y'' - 3xy' + 4y = 0, \quad \{x^2, x^2 \ln x\}, \quad$ on $x > 0$
(j) $y'' - 4y' + 4y = 0, \quad \{e^{2x}, xe^{2x}\}$
(j) $x^3y''' + xy' - y = 0, \quad \left\{x, x\ln x, x(\ln x)^2\right\}, \quad$ on $x > 0$

5. (*Liouville's formula*) (a) Derive Liouville's formula, (11), for the special case where $n = 2$, by writing out $W'(x)$, showing that

$$W'(x) = -p_1(x)W(x), \tag{5.1}$$

and integrating the latter to obtain (11).
(b) Derive (11) for the general case (i.e., where n need not equal 2), by showing that $W'(x)$ is the sum of n determinants where the jth one is obtained from the W determinant by differentiating the jth row and leaving the other rows unchanged. Show that each of these n determinants, except the nth one, has two identical rows and hence vanishes, so that

$$W'(x) = \begin{vmatrix} u_1(x) & \cdots & u_n(x) \\ \vdots & & \vdots \\ u_1^{(n-2)}(x) & \cdots & u_n^{(n-2)}(x) \\ u_1^{(n)}(x) & \cdots & u_n^{(n)}(x) \end{vmatrix}. \tag{5.2}$$

In the last row, substitute $u^{(n)}(x) = -p_1(x)u^{(n-1)}(x) - \cdots - p_n(x)u(x)$ from (10), again omit vanishing determinants, and again obtain (5.1) and hence the solution (11). HINT: You may use the various properties of determinants, given in Section 10.4.

6. (a) Prove Theorem 3.2.4.
(b) Prove Theorem 3.2.5.
(c) Prove Theorem 3.2.6.

7. If u_1 and u_2 are LI, u_1 and u_3 are LI, and u_2 and u_3 are LI, does it follow that $\{u_1, u_2, u_3\}$ is LI? Prove or disprove. HINT: If a proposition is false it can be disproved by a single counterexample, but if it is true then a single example does not suffice as proof.

8. Verify that x^2 and x^3 are solutions of $x^2y'' - 4xy' + 6y = 0$ on $-\infty < x < \infty$. Also verify, from Theorem 3.2.4, that they are LI on that interval. Does the fact that their Wronskian $W[x^2, x^3](x) = x^4$ vanishes at $x = 0$, together with their linear independence on $-\infty < x < \infty$ violate Theorem 3.2.3? Explain.

3.3 Homogeneous Equation: General Solution

3.3.1. General solution. We studied the first-order linear homogeneous equation

$$y' + p(x)y = 0 \tag{1}$$

in Chapter 2, where $p(x)$ is continuous on the x interval of interest, I, and found the solution to be

$$y(x) = Ce^{-\int p(x)\,dx}, \tag{2}$$

where C is an arbitrary constant. If we append to (1) an initial condition $y(a) = b$, where a is a point in I, then we obtain, from (2),

$$y(x) = be^{-\int_a^x p(\xi)\,d\xi}, \tag{3}$$

as was shown in Section 2.2.

The solution (2) is really a family of solutions because of the arbitrary constant C. We showed that (2) contains *all* solutions of (1), so we called it a general solution of (1). In contrast, (3) was only one member of that family, so we called it a particular solution.

Now we turn to the nth-order linear equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{4}$$

and once again we are interested in general and particular solutions. By a **general solution** of (4), on an interval I, we mean a family of solutions that contains every solution of (4) on that interval, and by a **particular solution** of (4), we mean any one member of that family of solutions.

We begin with a fundamental existence and uniqueness theorem.*

THEOREM 3.3.1 *Existence and Uniqueness for Initial-Value Problem*
If $p_1(x), \dots, p_n(x)$ are continuous on a closed interval I, then the initial-value problem consisting of the differential equation

$$\frac{d^n y}{dx^n} + p_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{dy}{dx} + p_n(x)y = 0, \tag{5a}$$

together with initial conditions

$$y(a) = b_1,\ y'(a) = b_2,\ \dots,\ y^{(n-1)}(a) = b_n, \tag{5b}$$

*For a more complete sequence of theorems, and their proofs, we refer the interested reader to the little book by J. C. Burkill, *The Theory of Ordinary Differential Equations* (Edinburgh: Oliver and Boyd, 1956) or to William E. Boyce and Richard C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 5th ed. (New York: Wiley, 1992).

where the initial point a is in I, has a solution on I, and that solution is unique.

Notice how the initial conditions listed in (5b) are perfect – not too few and not too many – in narrowing the general solution of (5a) down to a unique particular solution, for (5a) gives $y^{(n)}(x)$ as a linear combination of $y^{(n-1)}(x), \ldots, y(x)$, the derivative of (5a) gives $y^{(n+1)}(x)$ as a linear combination of $y^{(n)}(x), \ldots, y(x)$, and so on. Thus, knowing $y(a), \ldots, y^{(n-1)}(a)$ we can use the differential equation (5a) and its derivatives to compute $y^{(n)}(a)$, $y^{(n+1)}(a)$, and so on, and therefore to develop the Taylor series of $y(x)$ about the point a; that is, to determine $y(x)$.

Let us leave the initial-value problem (5) now, and turn our attention to determining the nature of the general solution of the nth-order linear homogeneous equation (4). We begin by re-expressing (4) in the compact form

$$L[y] = 0, \tag{6}$$

where

$$L = \frac{d^n}{dx^n} + p_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{d}{dx} + p_n(x) \tag{7}$$

is called an nth-order **differential operator** and

$$\begin{aligned} L[y] &= \left(\frac{d^n}{dx^n} + p_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + p_{n-1}(x)\frac{d}{dx} + p_n(x)\right)[y] \\ &\equiv \frac{d^n}{dx^n}y(x) + p_1(x)\frac{d^{n-1}}{dx^{n-1}}y(x) + \cdots + p_n(x)y(x) \end{aligned} \tag{8}$$

defines the action of L on any n-times differentiable function y. $L[y]$ is itself a function of x, with values $L[y](x)$. For instance, if $n = 2$, $p_1(x) = \sin x$, $p_2(x) = 5x$, and $y(x) = x^2$, then $L[y](x) = (x^2)'' + (\sin x)(x^2)' + 5x(x^2) = 2 + 2x\sin x + 5x^3$.

The key property of the operator L defined by (8) is that

$$\boxed{L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]} \tag{9}$$

for any (n-times differentiable) functions u, v and any constants α, β. Let us verify (9) for the representative case where L is of second order:

$$\begin{aligned} L[\alpha u + \beta v] &= \left(\frac{d^2}{dx^2} + p_1\frac{d}{dx} + p_2\right)(\alpha u + \beta v) \\ &= (\alpha u + \beta v)'' + p_1(\alpha u + \beta v)' + p_2(\alpha u + \beta v) \\ &= \alpha u'' + \beta v'' + p_1\alpha u' + p_1\beta v' + p_2\alpha u + p_2\beta v \\ &= \alpha\left(u'' + p_1 u' + p_2 u\right) + \beta\left(v'' + p_1 v' + p_2 v\right) \\ &= \alpha L[u] + \beta L[v]. \end{aligned} \tag{10}$$

Similarly for $n \geq 3$.

Recall that the differential equation (4) was classified as linear. Likewise, the corresponding operator L given by (8) is said to be a **linear differential operator**. The key and defining feature of a linear differential operator is the linearity property (9), which will be of great importance to us.

In fact, (9) holds not just for two functions u and v, but for any finite number of functions, say $u_1, \ldots, u_k$. That is,

$$\boxed{L\left[\alpha_1 u_1 + \cdots + \alpha_k u_k\right] = \alpha_1 L\left[u_1\right] + \cdots + \alpha_k L\left[u_k\right]} \tag{11}$$

for any functions $u_1, \ldots, u_k$ and any constants $\alpha_1, \ldots, \alpha_k$. (Of course it should be understood, whether we say so explicitly or not, that $u_1, \ldots, u_k$ must be n times differentiable since they are being operated on by the nth-order differential operator L.) To prove (11) we apply (9) step by step. For instance, if $k = 3$ we have

$$\begin{aligned} L\left[\alpha_1 u_1 + \alpha_2 u_2 + \alpha_3 u_3\right] &= L\left[\alpha_1 u_1 + 1\left(\alpha_2 u_2 + \alpha_3 u_3\right)\right] \\ &= \alpha_1 L\left[u_1\right] + 1 L\left[\alpha_2 u_2 + \alpha_3 u_3\right] && \text{from (9)} \\ &= \alpha_1 L\left[u_1\right] + \alpha_2 L\left[u_2\right] + \alpha_3 L\left[u_3\right] && \text{from (9) again.} \end{aligned}$$

From (11) we have the following superposition theorem:

THEOREM 3.3.2 *Superposition of Solutions of (4)*
If $y_1 \ldots, y_k$, are solutions of (4), then $C_1 y_1 + \cdots + C_k y_k$ is too, for any constants $C_1, \ldots, C_k$.

Proof: It follows from (11) that

$$\begin{aligned} L\left[C_1 y_1 + \cdots + C_k y_k\right] &= C_1 L\left[y_1\right] + \cdots + C_k L\left[y_k\right] \\ &= C_1(0) + \cdots + C_k(0). \quad \blacksquare \end{aligned}$$

EXAMPLE 1. *Superposition.* It is readily verified, by direct substitution, that $y_1 = e^{3x}$ and $y_2 = e^{-3x}$ are solutions of the equation

$$y'' - 9y = 0. \tag{12}$$

(We are not yet concerned with how to find such solutions; we will come to that in the next section.) Thus $y = C_1 e^{3x} + C_2 e^{-3x}$ is also a solution, as can be verified by substituting it into (12), for any constants C_1, C_2. ▮

To emphasize that the theorem does not hold for nonlinear or nonhomogeneous equations, we offer two counter-examples:

EXAMPLE 2. It can be verified that $y_1 = 1$ and $y_2 = x^2$ are solutions of the nonlinear equation $x^3y'' - yy' = 0$, yet their linear combination $4 + 3x^2$ is not. ▮

EXAMPLE 3. It can be verified that $y_1 = 4e^{3x} - 2$ and $y_2 = e^{3x} - 2$ are solutions of the nonhomogeneous equation $y'' - 9y = 18$, yet their sum $5e^{3x} - 4$ is not. ▮

We can now prove the following fundamental result.

THEOREM 3.3.3 *General Solution of (4)*
Let $p_1(x), \ldots, p_n(x)$ be continuous on an open interval I. Then the nth-order linear homogeneous differential equation (4) admits exactly n LI solutions; that is, at least n and no more than n. If $y_1(x), \ldots, y_n(x)$ is such a set of LI solutions on I, then the general solution of (4), on I, is

$$y(x) = C_1y_1(x) + \cdots + C_ny_n(x), \tag{13}$$

where $C_1, \ldots, C_n$ are arbitrary constants.

Proof: To show that (4) has no more than n LI solutions, suppose that $y_1(x), \ldots, y_m(x)$ are solutions of (4), where $m > n$. Let ξ be some point in I. The n linear algebraic equations

$$\begin{aligned} c_1y_1(\xi) + \cdots + c_my_m(\xi) &= 0 \\ &\vdots \\ c_1y_1^{(n-1)}(\xi) + \cdots + c_my_m^{(n-1)}(\xi) &= 0 \end{aligned} \tag{14}$$

in the m unknown c's have nontrivial solutions because $m > n$. Choosing such a nontrivial set of c's, define

$$v(x) \equiv c_1y_1(x) + \cdots + c_my_m(x), \tag{15}$$

and observe first that

$$\begin{aligned} L[v] &= L[c_1y_1 + \cdots + c_my_m] \\ &= c_1L[y_1] + \cdots + c_mL[y_m] = c_1(0) + \cdots + c_m(0) = 0, \end{aligned} \tag{16}$$

where L is the differential operator in (4) and, second, that $v(\xi) = v'(\xi) = \cdots = v^{(n-1)}(\xi) = 0$. One function $v(x)$ that satisfies $L[v] = 0$ and $v(\xi) = \cdots = v^{(n-1)}(\xi) = 0$ is simply $v(x) = 0$. By the uniqueness part of Theorem 3.3.1 it is the only such function, so $v(x) = 0$. Recalling that the c's in $v(x) = c_1y_1(x) + \cdots + c_my_m(x) = 0$ are not all zero, it follows that $y_1(x), \ldots, y_m(x)$ must be LD. Thus, (4) cannot have more than n LI solutions.

To show that there are indeed n LI solutions of (4), let us put forward n such solutions. According to the existence part of Theorem 3.3.1, there must be solutions $y_1(x), \dots, y_n(x)$ of (4) satisfying the initial conditions

$$\begin{array}{llll} y_1(a) = \alpha_{11}, & y_1'(a) = \alpha_{12}, & \cdots & y_1^{(n-1)}(a) = \alpha_{1n}, \\ \vdots & \vdots & & \vdots \\ y_n(a) = \alpha_{n1}, & y_n'(a) = \alpha_{n2}, & \cdots & y_n^{(n-1)}(a) = \alpha_{nn}, \end{array} \tag{17}$$

where a is any chosen point in I and the α's are any chosen numbers such that their determinant is nonzero. (For instance, one such set of α's is given by $\alpha_{ii} = 1$ for each $i = 1$ through n and $\alpha_{ij} = 0$ whenever $i \neq j$.) According to Theorem 3.2.3, $y_1(x), \dots, y_n(x)$ must be LI since their Wronskian is nonzero at $x = a$. Thus, there are indeed n LI solutions of (4).

Finally, every solution of (4) must be expressible as a linear combination of those n LI solutions, as in (13), for otherwise there would be more than n LI solutions of (4). ■

Any such set of n LI solutions is called a **basis**, or **fundamental set**, of solutions of the differential equation.

EXAMPLE 4. Suppose we begin writing solutions of the equation $y'' - 9y = 0$, from Example 1: $e^{3x}, 5e^{3x}, e^{-3x}, 2e^{3x} + e^{-3x}, \sinh 3x, \cosh 3x, e^{3x} - 4\cosh 3x$, and so on. (That each is a solution is easily verified.) From among these we can indeed find two that are LI, but no more than two. For instance, $\{e^{3x}, e^{-3x}\}$, $\{e^{3x}, 2e^{3x} + e^{-3x}\}$, $\{e^{3x}, \sinh 3x\}$, $\{\sinh 3x, \cosh 3x\}$, $\{\sinh 3x, e^{-3x}\}$ are bases, so the general solution can be expressed in any of these ways:

$$y(x) = C_1 e^{3x} + C_2 e^{-3x}, \tag{18a}$$
$$y(x) = C_1 e^{3x} + C_2 \left(2e^{3x} + e^{-3x}\right), \tag{18b}$$
$$y(x) = C_1 e^{3x} + C_2 \sinh 3x, \tag{18c}$$
$$y(x) = C_1 \sinh 3x + C_2 \cosh 3x, \tag{18d}$$

and so on. Each of these is a general solution of $y'' - 9y = 0$, and all of them are equivalent. For instance, (18a) implies (18d) because

$$\begin{aligned} y(x) &= C_1 e^{3x} + C_2 e^{-3x} \\ &= C_1 (\cosh 3x + \sinh 3x) + C_2 (\cosh 3x - \sinh 3x) \\ &= (C_1 + C_2)\cosh 3x + (C_1 - C_2)\sinh 3x \\ &= C_1' \cosh 3x + C_2' \sinh 3x, \end{aligned}$$

where C_1', C_2' are arbitrary constants (Exercise 15). ■

EXAMPLE 5. Solve the initial-value problem

$$y''' + y' = 0 \tag{19a}$$

$$y(0) = 3, \quad y'(0) = 5, \quad y''(0) = -4. \tag{19b}$$

A general solution of (19a) is

$$y(x) = C_1 \cos x + C_2 \sin x + C_3, \tag{20}$$

because $\cos x$, $\sin x$, and 1 are LI solutions of (19a). Imposing (19b) gives

$$\begin{aligned} y(0) &= 3 = C_1 + C_3, \\ y'(0) &= 5 = C_2, \\ y''(0) &= -4 = -C_1, \end{aligned}$$

which equations admit the unique solution $C_1 = 4, C_2 = 5, C_3 = -1$. Thus,

$$y(x) = 4 \cos x + 5 \sin x - 1$$

is the unique solution to the initial-value problem (19). ▮

3.3.2. Boundary-value problems. It must be remembered that the existence and uniqueness theorem, Theorem 3.3.1, is for *initial*-value problems. Though most of our interest is in initial-value problems, one also encounters problems of boundary-value type, where conditions are specified at *two* points, normally the ends of the interval I. Not only are boundary-value problems not covered by Theorem 3.3.1, it is striking that *boundary-value problems need not have unique solutions.* In fact, they may have *no solution, a unique solution, or a nonunique solution*, as shown by the following example.

EXAMPLE 6. *Boundary-Value Problem.* It is readily verified that the differential equation

$$y'' + y = 0 \tag{21}$$

admits a general solution

$$y(x) = C_1 \cos x + C_2 \sin x. \tag{22}$$

Consider three different sets of boundary values.

Case 1: $y(0) = 2, y(\pi) = 1$. Then

$$\begin{aligned} y(0) &= 2 = C_1 + 0, \\ y(\pi) &= 1 = -C_1 + 0, \end{aligned}$$

which has no solution for C_1, C_2, so the boundary-value problem has no solution for $y(x)$.

Case 2: $y(0) = 2, y(\pi/2) = 3$. Then

$$\begin{aligned} y(0) &= 2 = C_1 + 0, \\ y(\pi/2) &= 3 = 0 + C_2, \end{aligned}$$

so $C_1 = 2, C_2 = 3$, and the boundary-value problem has the unique solution $y(x) = 2\cos x + 3\sin x$.

Case 3: $y(0) = 2$, $y(\pi) = -2$. Then

$$\begin{aligned} y(0) &= 2 = C_1 + 0, \\ y(\pi) &= -2 = -C_1 + 0, \end{aligned}$$

so $C_1 = 2$, and C_2 is arbitrary, and the boundary-value problem has the nonunique solution (indeed, the infinity of solutions) $y(x) = 2\cos x + C_2 \sin x$, where C_2 is an arbitrary constant. ∎

Boundary-value problems are studied further in Sections 6.4, 17.7, and 17.8.

Closure. In this section we have considered the nth-order linear homogeneous differential equation $L[y] = 0$. The principal result was that a general solution

$$y(x) = C_1 y_1(x) + \cdots + C_n y_n(x)$$

can be built up by the superposition of n LI solutions $y_1(x), \ldots, y_n(x)$, thanks to the linearity of L, namely, the property of L that

$$L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]$$

for any n-times differentiable functions u, v and any constants α, β. Any such set of n LI solutions $\{y_1, \ldots, y_k\}$ of $L[y] = 0$ is called a **basis of solutions** (or **basis**, for brevity) for that equation.

For the initial-value problem consisting of the differential equation $L[y] = 0$ together with the initial conditions (5b), we found that a solution exists and is unique, but for the boundary-value version we found that a solution need not exist, it may exist and be unique, or there may exist a nonunique solution.

Theorems 3.3.1 and 3.3.3 are especially important.

EXERCISES 3.3

NOTE: If not specified, the interval I is understood to be the entire x axis.

1. Show whether or not each of the following is a general solution to the equation given.

(a) $y'' - 3y' + 2y = 0$; $\quad C_1 e^x + C_2 e^{2x}$
(b) $y'' - 3y' + 2y = 0$; $\quad C_1(e^x - e^{2x}) + C_2 e^x$
(c) $y'' - y' - 2y = 0$; $\quad C_1(e^{-x} + e^{2x})$
(d) $y'' - y' - 2y = 0$; $\quad C_1 e^{-x} + C_2 e^{2x}$
(e) $y''' + 4y' = 0$; $\quad C_1 \cos 2x + C_2 \sin 2x$
(f) $y''' + 4y' = 0$; $\quad C_1 + C_2 \cos 2x + C_2 \sin 2x$
(g) $y''' - 2y'' + y' = 0$; $\quad \left(C_1 + C_2 x + C_3 x^2\right) e^x$
(h) $y''' - 2y'' + y' = 0$; $\quad (C_1 + C_2 x) e^x + C_3$
(i) $y''' - y'' - y' + y = 0$; $\quad C_1 e^x + C_2 e^{-x} + C_3 x e^x$

2. Show whether or not each of the following is a basis for the given equation.

(a) $y'' - 9y = 0$; $\quad \left\{e^{3x}, \cosh 3x, \sinh 3x\right\}$
(b) $y'' - 9y = 0$; $\quad \left\{e^{3x}, \cosh 3x\right\}$
(c) $y'' - y = 0$; $\quad \{\sinh 3x, 2\cosh 3x\}$
(d) $y''' - 3y'' + 3y' - y = 0$; $\quad \left\{e^x, xe^x, x^2 e^x\right\}$
(e) $y''' - 3y'' = 0$; $\quad \left\{1, x, e^{3x}\right\}$

(f) $y'''' + 2y'' + y = 0$; $\{\cos x, \sin x, x\cos x, x\sin x\}$

3. Are the following general solutions of $x^2y'' + xy' - 4y = 0$ on $0 < x < \infty$? On $-\infty < x < \infty$? Explain.

(a) C_1x^2 (b) $C_1x^2 + C_2x^{-2}$
(c) $C_1(x^2 + x^{-2}) + C_2(x^2 - x^{-2})$

4. Are the following bases for the equation $x^2y'' - xy' + y = 0$ on $0 < x < \infty$? On $-\infty < x < 0$? On $-\infty < x < \infty$? On $6 < x < 10$? Explain.

(a) $\{x, x^2\}$
(b) $\{e^x, e^{-x}\}$
(c) $\{x, x\ln|x|\}$
(d) $\{x + x\ln|x|, x - x\ln|x|\}$

5. Show whether or not the following is a general solution of $y^{(vii)} - 4y^{(vi)} - 14y^{(v)} + 56y^{(iv)} + 49y''' - 196y'' - 36y' + 144y = 0$.

(a) $C_1e^x + C_2e^{-x} + C_3e^{2x} + C_4e^{-2x} + C_5e^{3x} + C_6e^{-3x}$
(b) $C_1e^x + C_2e^{-x} + C_3e^{2x} + C_4e^{-2x} + C_5e^{3x} + C_6e^{-3x} + C_7\sinh x + C_8\cosh 2x$

6. Show that $y_1 = 1$ and $y_2 = 2$ are solutions of $(y^3 - 6y^2 + 11y - 6)\,y'' = 0$. Is $y = y_1 + y_2 = 1 + 2 = 3$ a solution as well? Does your result contradict the sentence preceding Example 2? Explain.

7. Show that each of the functions $y_1 = 3x^2 - x$ and $y_2 = x^2 - x$ is a solution of the equation $x^2y'' - 2y = 2x$. Is the linear combination $C_1y_1 + C_2y_2$ a solution as well, for all choices of the constants C_1 and C_2?

8. (*Taylor series method*) Use the Taylor series method described below Theorem 3.3.1 to solve each of the following initial-value problems for $y(x)$, up to and including terms of fifth order. NOTE: The term $f^{(n)}(a)(x-a)^n/n!$ in the Taylor series of $f(x)$ about $x = a$ is said to be of ***n*th-order**.

(a) $y'' + y = 0$; $y(0) = 4,\ y'(0) = 3$
(b) $y'' - 4y = 0$; $y(0) = -1,\ y'(0) = 0$
(c) $y'' + 5y' + 6y = 0$; $y(0) = 2,\ y'(0) = -5$
(d) $y'' + xy = 0$; $y(0) = 1,\ y'(0) = 0$
(e) $y'' + x^2y = 0$; $y(0) = 2,\ y'(0) = -3$
(f) $y'' - 3y = 0$; $y(5) = 4,\ y'(5) = 6$ HINT: Expand about $x = 5$.
(g) $y'' + 3y' - y = 0$; $y(1) = 2,\ y'(1) = 0$ HINT: Expand about $x = 1$.
(h) $y''' - y' + 2y = 0$; $y(0) = 0,\ y'(0) = 0,\ y''(0) = 1$
(i) $y''' - xy = 0$; $y(0) = 0,\ y'(0) = 3,\ y''(0) = -2$

9. Does the problem stated have a unique solution? No solution? A nonunique solution? Explain.

(a) $y'' + 2y' + 3y = 0$; $y(0) = 5,\ y'(0) = -1$
(b) $y'' + 2y' + 3y = 0$; $y(3) = 2,\ y'(3) = 37$
(c) $y'' + xy' - y = 0$; $y(3) = y'(3) = 0$
(d) $xy''' + xy' - y = 0$; $y(-1) = y'(-1) = 0,\ y''(-1) = 4$
(e) $x^2y'' - y' - y = 0$; $y(6) = 0,\ y'(6) = 1$
(f) $(\sin x)y'''' + xy''' = 0$; $y(2) = y'(2) = y''(2) = 0$, $y'''(2) = -9$

10. Verify that (22) is indeed a general solution of (21).

11. Consider the boundary-value problem consisting of the differential equation $y'' + y = 0$ plus the boundary conditions given. Does the problem have any solutions? If so, find them. Is the solution unique? HINT: A general solution of the differential equation is $y = C_1\cos x + C_2\sin x$.

(a) $y(0) = 0,\ y(2) = 0$
(b) $y(0) = 0,\ y(2\pi) = -3$
(c) $y(1) = 1,\ y(2) = 2$
(d) $y'(0) = 0,\ y(5) = 1$
(e) $y'(0) = 0,\ y'(\pi) = 0$
(f) $y'(0) = 0,\ y'(6\pi) = 0$
(g) $y'(0) = 0,\ y'(2\pi) = 38$

12. Consider the boundary-value problem consisting of the differential equation $y'''' + 2y'' + y = 0$ plus the boundary conditions given. Does the problem have any solutions? If so, find them. Is the solution unique? HINT: A general solution of the differential equation is $y = (C_1 + C_2x)\cos x + (C_3 + C_4x)\sin x$.

(a) $y(0) = y'(0) = 0,\ y(\pi) = 0,\ y'(\pi) = 2$
(b) $y(0) = y'(0) = y''(0) = 0,\ y(\pi) = 1$
(c) $y(0) = y''(0) = 0,\ y(\pi) = 0 = y''(\pi) = 0$
(d) $y(0) = y''(0) = 0,\ y(\pi) = y''(\pi) = 3$

13. Prove that the linearity property (10) is equivalent to the two properties

$$L[u+v] = L[u] + L[v], \tag{13.1a}$$

$$L[\alpha u] = \alpha L[u]. \tag{13.1b}$$

That is, show that the truth of (10) implies the truth of (13.1), and conversely.

14. We showed that (11) holds for the case $k = 3$, but did not prove it in general. Here, we ask you to prove (11) for any integer $k \geq 1$. HINT: It is suggested that you use **mathematical induction**, whereby a proposition $P(k)$, for $k \geq 1$, is proved by first showing that it holds for $k = 1$, and then showing that if it holds for k then it must also hold for $k + 1$. In the present example, the proposition $P(k)$ is the equation (11).

15. (*Example 4, Continued*) (a) Verify that each of (18a)

through (18d) is a general solution of $y'' - 9y = 0$.
(b) It seems reasonable that if C_1, C_2 are arbitrary constants, and if we call

$$C_1 + C_2 = C_1' \quad \text{and} \quad C_1 - C_2 = C_2', \qquad (15.1)$$

then C_1', C_2' are arbitrary too, as we claimed at the end of Example 4. Actually, for that claim to be true we need to be able to show that corresponding to any chosen values C_1', C_2' the equations (15.1) on C_1, C_2 are consistent – that is, that they admit one or more solutions for C_1, C_2. Show that (15.1) is indeed consistent.
(c) Show that if, instead, we had $C_1 + C_2 = C_1'$ and $2C_1 + 2C_2 = C_2'$, where C_1, C_2 are arbitrary constants, then it is *not* true that C_1', C_2' are arbitrary too.

3.4 Solution of Homogeneous Equation: Constant Coefficients

Knowing that the general solution of an nth-order homogeneous differential equation is an arbitrary linear combination of n LI (linearly independent) solutions, the question is: How do we find those solutions? That question will occupy us for the remainder of this chapter and for the next three chapters as well. In this section we consider the constant-coefficient case,

$$\frac{d^n y}{dx^n} + a_1 \frac{d^{n-1} y}{dx^{n-1}} + \cdots + a_{n-1} \frac{dy}{dx} + a_n y = 0; \qquad (1)$$

that is, where the a_j coefficients are constants, not functions of x. This case is said to be "elementary" in the sense that the solutions will always be found among the *elementary functions* (powers of x, trigonometric functions, exponentials, and logarithms), but it is also elementary in the sense that it is the simplest case: nonconstant-coefficient equations are generally much harder, and nonlinear equations are much harder still.

Fortunately, the constant-coefficient case is not only the simplest, it is also of great importance in science and engineering. For instance, the equations

$$mx'' + cx' + kx = 0$$

and

$$Li'' + Ri' + \frac{1}{C} i = 0,$$

governing mechanical and electrical oscillators, where primes denote derivatives with respect to the independent variable t, are both of the type (1) because m, c, k and L, R, C are constants; they do not vary with t.

3.4.1. Euler's formula and review of the circular and hyperbolic functions. We are going to be involved with the exponential function e^z, where $z = x + iy$ is complex and $i = \sqrt{-1}$. The first point to appreciate is that we cannot figure out how to evaluate e^{x+iy} from our knowledge of the function e^x where x is real. That is, e^{x+iy} is a "new object," and its values are a matter of definition, not a matter of figuring out. To motivate that definition, let us proceed as follows:

$$\begin{aligned} e^z = e^{x+iy} &= e^x e^{iy} \\ &= e^x \left[1 + iy + \frac{(iy)^2}{2!} + \frac{(iy)^3}{3!} + \frac{(iy)^4}{4!} + \cdots \right] \\ &= e^x \left[\left(1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \cdots \right) + i \left(y - \frac{y^3}{3!} + \frac{y^5}{5!} - \cdots \right) \right]. \end{aligned} \tag{2}$$

Recognizing the two series as the Taylor series representations of $\cos y$ and $\sin y$, respectively, we obtain

$$\boxed{e^{x+iy} = e^x \left(\cos y + i \sin y \right),} \tag{3}$$

which is known as **Euler's formula**, after the great Swiss mathematician *Leonhard Euler* (1707–1783), whose many contributions to mathematics included the systematic development of the theory of linear constant-coefficient differential equations.

We say that (3) defines e^{x+iy} since it gives e^{x+iy} in the standard Cartesian form $a + ib$, where the real part a is $e^x \cos y$ and the imaginary part b is $e^x \sin y$. Observe carefully that we cannot defend certain steps in (2). Specifically, the second equality seems to be the familiar formula $e^{a+b} = e^a e^b$, but the latter is for real numbers a and b, whereas iy is not real. Likewise, the third equality rests upon the Taylor series formula $e^u = 1 + u + \frac{u^2}{2!} + \cdots$ that is derived in the calculus for the case where u is real, but iy is not real. The point to understand, then, is that the steps in (2) are merely heuristic; trying to stay as close to real-variable theory as possible, we arrive at (3). Once (3) is obtained, we throw out (2) and take (3) as our (i.e., Euler's) *definition* of e^{x+iy}. Of course, there are an infinite number of ways one can define a given quantity, but some are more fruitful than others. To fully appreciate why Euler's definition is the perfect one for e^{x+iy}, one needs to study complex-variable theory, as we will in later chapters. For the present, we merely propose that the steps in (2) make (3) a *reasonable* choice as a definition of e^z.

As a special case of (3), let $x = 0$. Then (3) becomes

$$e^{iy} = \cos y + i \sin y. \tag{4a}$$

For instance, $e^{\pi i} = \cos \pi + i \sin \pi = -1 + 0i = -1$, and $e^{2-3i} = e^2 (\cos 3 - i \sin 3) = 7.39(-0.990 - 0.141i) = -7.32 - 1.04i$. Since (4a) holds for all y, it must hold also with y changed to $-y$:

$$e^{-iy} = \cos(-y) + i \sin(-y),$$

and since $\cos(-y) = \cos y$ and $\sin(-y) = -\sin y$, it follows that

$$e^{-iy} = \cos y - i \sin y. \tag{4b}$$

Conversely, we can express $\cos y$ and $\sin y$ as linear combinations of the complex exponentials e^{iy} and e^{-iy}, for adding (4a) and (4b) and subtracting them gives $\cos y = \left(e^{iy} + e^{-iy}\right)/2$ and $\sin y = \left(e^{iy} - e^{-iy}\right)/(2i)$. Let us frame these formulas, for emphasis and reference:

$$\boxed{\begin{aligned} e^{iy} &= \cos y + i\sin y \\ e^{-iy} &= \cos y - i\sin y \end{aligned}} \tag{5a,b}$$

and

$$\boxed{\begin{aligned} \cos y &= \frac{e^{iy} + e^{-iy}}{2} \\ \sin y &= \frac{e^{iy} - e^{-iy}}{2i}. \end{aligned}} \tag{6a,b}$$

Observe that all four of these formulas come from the single formula (4a). (Of course there is nothing essential about the name of the variable in these formulas. For instance, $e^{ix} = \cos x + i\sin x$, $e^{i\theta} = \cos\theta + i\sin\theta$, and so on.)

There is a similarity between (5) and (6), relating the cosine and sine to the complex exponentials, to analogous formulas relating the hyperbolic cosine and hyperbolic sine to real exponentials. If we recall the definitions

$$\boxed{\begin{aligned} \cosh y &= \frac{e^{y} + e^{-y}}{2}, \\ \sinh y &= \frac{e^{y} - e^{-y}}{2} \end{aligned}} \tag{7a,b}$$

of the hyperbolic cosine and hyperbolic sine, we find, by addition and subtraction of these formulas, that

$$\boxed{\begin{aligned} e^{y} &= \cosh y + \sinh y, \\ e^{-y} &= \cosh y - \sinh y. \end{aligned}} \tag{8a,b}$$

Compare (5) with (8), and (6) with (7). The graphs of $\cosh x$, $\sinh x$, e^x, and e^{-x} are given in Fig. 1.

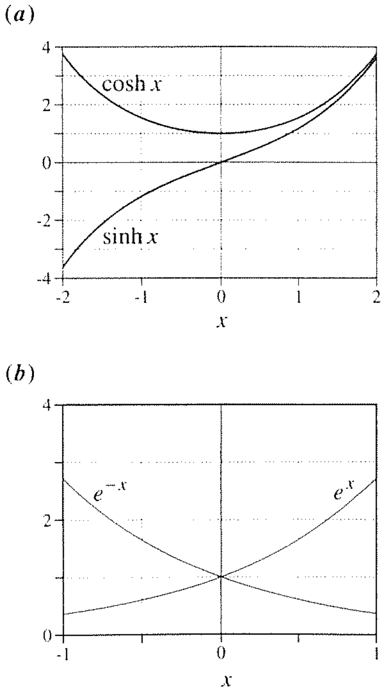


Figure 1. $\cosh x$, $\sinh x$, e^x, and e^{-x}.

Using (6) and (7) we obtain the properties

$$\cos^2 y + \sin^2 y = 1, \tag{9}$$

$$\cosh^2 y - \sinh^2 y = 1. \tag{10}$$

From a geometric point of view, if we parametrize a curve C by the relations

$$x = \cos\tau, \quad y = \sin\tau \tag{11}$$

over $0 \leq \tau < 2\pi$, say, then it follows from (9) that $x^2 + y^2 = 1$, so that C is a circle. And if we parametrize C by

$$x = \cosh\tau, \quad y = \sinh\tau \tag{12}$$

instead, then it follows from (10) that $x^2 - y^2 = 1$, so C is a hyperbola. Thus, one refers to $\cos x$ and $\sin x$ as *circular functions* and to $\cosh x$ and $\sinh x$ as *hyperbolic functions*, the *hyperbolic cosine* and the *hyperbolic sine*, respectively.

Besides (9) and (10), various useful identities, such as

$$\sin(A+B) = \sin A\cos B + \sin B\cos A, \tag{13a}$$
$$\cos(A+B) = \cos A\cos B - \sin A\sin B, \tag{13b}$$
$$\sinh(A+B) = \sinh A\cosh B + \sinh B\cosh A, \tag{13c}$$
$$\cosh(A+B) = \cosh A\cosh B + \sinh A\sinh B, \tag{13d}$$

can be derived from (6) and (7), as well as the derivative formulas

$$\begin{aligned}\frac{d}{dx}\cos x &= -\sin x, &\quad \frac{d}{dx}\sin x &= \cos x,\\ \frac{d}{dx}\cosh x &= \sinh x, &\quad \frac{d}{dx}\sinh x &= \cosh x.\end{aligned} \tag{14}$$

We shall be interested specifically in the function $e^{\lambda x}$ and its derivatives with respect to x, where λ is a constant that may be complex, say $\lambda = a + ib$. We know from the calculus that

$$\frac{d}{dx}e^{\lambda x} = \lambda e^{\lambda x} \tag{15}$$

when λ is a real constant. Does (15) hold when λ is complex? To answer the question, use Euler's formula (3) to express

$$e^{\lambda x} = e^{(a+ib)x} = e^{ax}(\cos bx + i\sin bx).$$

Thus,

$$\begin{aligned}\frac{d}{dx}e^{\lambda x} &= \frac{d}{dx}\left[e^{ax}(\cos bx + i\sin bx)\right]\\ &= \frac{d}{dx}(e^{ax}\cos bx) + i\frac{d}{dx}(e^{ax}\sin bx)\\ &= (ae^{ax}\cos bx - be^{ax}\sin bx) + i(ae^{ax}\sin bx + be^{ax}\cos bx)\\ &= e^{ax}(a+ib)(\cos bx + i\sin bx)\\ &= \lambda e^{ax}(\cos bx + i\sin bx) = \lambda e^{\lambda x},\end{aligned}$$

so the familiar formula (15) does hold even for complex λ.

There is one more fact about the exponential function that we will be needing, namely, that the exponential function e^z cannot be zero for any choice of z; that is, *it has no zeros*, for

$$\begin{aligned}|e^z| &= \left|e^{x+iy}\right| = |e^x(\cos y + i\sin y)|\\ &= |e^x|\,|\cos y + i\sin y| = e^x\,|\cos y + i\sin y| = e^x.\end{aligned}$$

The fourth equality follows from the fact that the real exponential is everywhere positive, and the fifth equality from the fact that $|a + ib|$ is the square root of the sum of the squares of a and b, and $\cos^2 y + \sin^2 y = 1$. Finally, we know that $e^x > 0$ for all x, so $|e^z| > 0$ for all z, and hence $e^z \neq 0$ for all z, as claimed.

3.4.2. Exponential solutions. To guide our search for solutions of (1), it is a good idea to begin with the simplest case, $n = 1$:

$$\frac{dy}{dx} + a_1 y = 0, \tag{16}$$

the general solution of which is

$$y(x) = Ce^{-a_1 x}, \tag{17}$$

where C is an arbitrary constant. One can derive (17) by noticing that (16) is a first-order linear equation and using the general solution developed in Section 2.2, or by using the fact that (16) is separable.

Observing that (17) is of exponential form, it is natural to wonder if higher-order equations admit exponential solutions too. Consider the second-order equation

$$y'' + a_1 y' + a_2 y = 0, \tag{18}$$

where a_1 and a_2 are real numbers, and let us seek a solution in the form

$$y(x) = e^{\lambda x}. \tag{19}$$

If (19) is to be a solution of (18), then it must be true that

$$\lambda^2 e^{\lambda x} + a_1 \lambda e^{\lambda x} + a_2 e^{\lambda x} = 0, \tag{20}$$

or

$$\left(\lambda^2 + a_1 \lambda + a_2\right) e^{\lambda x} = 0, \tag{21}$$

where (20) holds, according to (15), even if the not-yet-determined constant λ turns out to be complex. For (19) to be a solution of (18) on some interval I, we need (21) to be satisfied on I. That is, we need the left side of (21) to vanish identically on I. Since $e^{\lambda x}$ is not identically zero on any interval I for any choice of λ, we need λ to be such that

$$\lambda^2 + a_1 \lambda + a_2 = 0. \tag{22}$$

This equation and its left-hand side are called the **characteristic equation** and **characteristic polynomial**, respectively, corresponding to the differential equation (18). In general, (22) gives two distinct roots, say λ_1 and λ_2, which can be found from the quadratic formula as

$$\lambda = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2}}{2}.$$

(The nongeneric case of repeated roots, which occurs if $a_1^2 - 4a_2$ vanishes, is discussed separately, below.)

Thus, our choice of the exponential form (19) has been successful. Indeed, we have found *two* solutions of that form, $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$. Next, from Theorem 3.3.2 it follows [thanks to the linearity of (18)] that if $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$ are solutions of (18) then so is any linear combination of them,

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}. \tag{23}$$

Theorem 3.3.3 guarantees that (23) is a general solution of (18) if $e^{\lambda_1 x}$ and $e^{\lambda_2 x}$ are LI on I, and Theorem 3.2.4 tells us that they are indeed LI since neither one is expressible as a scalar multiple of the other. Thus, by seeking solutions in the form (19) we were successful in finding the general solution (23) of (18).

EXAMPLE 1. For the equation

$$y'' - y' - 6y = 0, \tag{24}$$

the characteristic equation is $\lambda^2 - \lambda - 6 = 0$, with roots $\lambda = -2, 3$, so

$$y(x) = C_1 e^{-2x} + C_2 e^{3x} \tag{25}$$

is a general solution of (24). ▮

EXAMPLE 2. For the equation

$$y'' - 9y = 0, \tag{26}$$

the characteristic equation is $\lambda^2 - 9 = 0$, with roots $\lambda = \pm 3$, so a general solution of (26) is

$$y(x) = C_1 e^{3x} + C_2 e^{-3x}. \tag{27}$$

COMMENT 1. As discussed in Example 4 of Section 3.3, an infinite number of forms of the general solution to (26) are equivalent to (27), such as

$$y(x) = C_1 \cosh 3x + C_2 \sinh 3x, \tag{28}$$

$$y(x) = C_1 \sinh 3x + C_2 \left(5e^{-3x} - 2\cosh 3x\right), \tag{29}$$

$$y(x) = C_1 \left(e^{3x} + 4\sinh 3x\right) + C_2 \left(\cosh 3x - \sqrt{\pi}\sinh 3x\right), \tag{30}$$

and so on. Of these, one would normally choose either (27) or (28). What is wrong with (29) and (30)? Nothing, except that they are ugly; e^{3x} and e^{-3x} make a "handsome couple," and $\cosh 3x$ and $\sinh 3x$ do too, but the choices in (29) and (30) seem ugly and purposeless.

COMMENT 2. If (27) and (28) are equivalent, does it matter whether we choose one or the other? No, since they are equivalent. However, one may be more convenient than the other insofar as the application of initial or boundary conditions.

For instance, suppose we append to (26) the initial conditions $y(0) = 4$, $y'(0) = -5$. Applying these to (27) gives

$$\begin{aligned} y(0) &= 4 = C_1 + C_2, \\ y'(0) &= -5 = 3C_1 - 3C_2, \end{aligned} \tag{31}$$

so $C_1 = 7/6$, $C_2 = 17/6$, and $y(x) = (7e^{3x} + 17e^{-3x})/6$. Applying these initial conditions to (28), instead, gives

$$\begin{aligned} y(0) &= 4 = C_1, \\ y'(0) &= -5 = 3C_2, \end{aligned} \tag{32}$$

so $C_1 = 4$, $C_2 = -5/3$, and $y(x) = 4\cosh 3x - (5/3)\sinh 3x$. Whereas our final results are equivalent, we see that (32) was more readily solved than (31). Thus, $\cosh 3x$ and $\sinh 3x$ make a slightly better choice in this case than e^{3x} and e^{-3x} – namely, when initial conditions are given at $x = 0$.

Or, suppose we consider I to be $0 \le x < \infty$ and impose the boundary conditions that $y(0) = 6$, and that $y(x)$ is to be bounded as $x \to \infty$. That is, rather than impose a numerical value on y at infinity, we impose a **boundedness condition**, that $|y(x)| < M$ for all x, for some constant M. Applying these conditions to (27) we see, from the boundedness condition, that we need $C_1 = 0$ since otherwise the e^{3x} will give unbounded growth. Next, $y(0) = 6 = C_2$, and hence the solution is $y(x) = 6e^{-3x}$. Notice how easily the boundedness condition was applied to (27).

If we use (28) instead, the solution is harder since both $\cosh 3x$ and $\sinh 3x$ grow unboundedly as $x \to \infty$. We can't afford to set both $C_1 = 0$ and $C_2 = 0$, in (28) since then we would have $y(x) = 0$, which does indeed satisfy both the equation (26) and the boundedness condition, but cannot satisfy the remaining initial condition $y(0) = 6$. However, perhaps the growth in $\cosh 3x$ and $\sinh 3x$ can be made to cancel. That is, write

$$\begin{aligned} y(x) &= C_1 \cosh 3x + C_2 \sinh 3x \\ &= C_1 \left(\frac{e^{3x} + e^{-3x}}{2} \right) + C_2 \left(\frac{e^{3x} - e^{-3x}}{2} \right) \\ &= \frac{C_1 + C_2}{2} e^{3x} + \frac{C_1 - C_2}{2} e^{-3x}, \end{aligned} \tag{33}$$

so for boundedness we need $C_1 + C_2 = 0$ (and hence $C_2 = -C_1$). Then (33) gives $y(x) = C_1 e^{-3x}$ and $y(0) = 6$ gives $C_1 = 6$ and $y(x) = 6e^{-3x}$, as before. Thus, in the case of a boundedness boundary condition at infinity we see that the exponential form (27) is more convenient than the hyperbolic form (28).

To summarize, when confronted with a choice, such as between (27) and (28), look ahead to the application of any initial or boundary conditions to see if one form will be more convenient than the other. ▮

EXAMPLE 3. For

$$y'' + 9y = 0, \tag{34}$$

the characteristic equation is $\lambda^2 + 9 = 0$, with roots $\lambda = \pm 3i$, so a general solution of (34) is

$$y(x) = C_1 e^{i3x} + C_2 e^{-i3x}. \tag{35}$$

COMMENT 1. Just as the general solution of $y'' - 9y = 0$ was expressible in terms of the real exponentials e^{3x}, e^{-3x} or the hyperbolic functions $\cosh 3x, \sinh 3x$, the general solution of (34) is expressible in terms of the complex exponentials e^{i3x}, e^{-i3x} or in terms of the circular functions $\cos 3x, \sin 3x$, for we can use Euler's formula to re-express (35) as

$$\begin{aligned} y(x) &= C_1(\cos 3x + i \sin 3x) + C_2(\cos 3x - i \sin 3x) \\ &= (C_1 + C_2)\cos 3x + i(C_1 - C_2)\sin 3x. \end{aligned} \tag{36}$$

Since C_1 and C_2 are arbitrary constants, we can simplify this result by letting $C_1 + C_2$ be a new constant A, and letting $i(C_1 - C_2)$ be a new constant B, so we have, from (36), the form

$$y(x) = A \cos 3x + B \sin 3x, \tag{37}$$

where A, B are arbitrary constants. As in Example 1, we note that (35) and (37) are but two out of an infinite number of equivalent forms.

COMMENT 2. You may be concerned that if $y(x)$ is a physical quantity such as the displacement of a mass or the current in an electrical circuit, then it should be real, whereas the right side of (35) seems to be complex. To explore this point, let us solve a complete problem, the differential equation (34) plus a representative set of initial conditions, say $y(0) = 7, y'(0) = 3$, and see if the final answer is real or not. Imposing the initial conditions on (35),

$$\begin{aligned} y(0) &= 7 = C_1 + C_2, \\ y'(0) &= 3 = i3C_1 - i3C_2, \end{aligned}$$

so $C_1 = (7-i)/2$ and $C_2 = (7+i)/2$. Putting these values into (35), we see from (36) that $y(x) = \frac{1}{2}[(7-i)+(7+i)]\cos 3x + \frac{1}{2}i[(7-i)-(7+i)]\sin 3x = 7\cos 3x + \sin 3x$, which is indeed real. Put differently, if the differential equation and initial conditions represent some physical system, then the mathematics "knows all about" the physics; it is built in, and we need not be anxious. ∎

Having already made the point that the general solution can always be expressed in various different (but equivalent) forms, we will generally adopt the exponential form when the exponentials are real, and the circular function form when they are complex. This decision is one of personal preference.

EXAMPLE 4. The equation

$$y'' + 4y' + 7y = 0 \tag{38}$$

has the characteristic equation $\lambda^2 + 4\lambda + 7 = 0$, with distinct roots $\lambda = -2 \pm i\sqrt{3}$, so a general solution of (38) is

$$\begin{aligned} y(x) &= C_1 e^{(-2+i\sqrt{3})x} + C_2 e^{(-2-i\sqrt{3})x} \\ &= e^{-2x}\left(C_1 e^{i\sqrt{3}x} + C_2 e^{-i\sqrt{3}x}\right) \\ &= e^{-2x}\left(A \cos\sqrt{3}x + B \sin\sqrt{3}x\right). \end{aligned} \tag{39}$$

That is, first we factor out the common factor e^{-2x}, then we re-express the complex exponentials in terms of the circular functions.

If we impose initial conditions $y(0) = 1, y'(0) = 0$, say, we find that $A = 1$ and $B = 2/\sqrt{3}$, so $y(x) = e^{-2x}\left(\cos\sqrt{3}x + \frac{2}{\sqrt{3}}\sin\sqrt{3}x\right)$. According to Theorem 3.3.1, that solution is unique. ▮

3.4.3. Higher-order equations $(n > 2)$**.** Examples 1–4 are representative of the four possible cases for second-order equations having distinct roots of the characteristic equation: if the roots are both real then the solution is expressible as a linear combination of two real exponentials (Example 1); if they are both real and equal and opposite in sign, then the solution is expressible either as exponentials or as a hyperbolic cosine and a hyperbolic sine (Example 2); if they are not both real then they will be complex conjugates. If those complex conjugates are purely imaginary, then the solution is expressible as a linear combination of two complex exponentials or as a sine and a cosine (Example 3); if they are not purely imaginary, then the solution is expressible as a real exponential times a linear combination of complex exponentials or a sine and a cosine (Example 4).

Turning to higher-order equations $(n > 2)$, our attention focuses on the characteristic equation

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0. \tag{40}$$

If $n = 1$, then (40) becomes $\lambda + a_1 = 0$ which, of course, has the root $\lambda = -a_1$ on the real axis. If $n = 2$, then (40) becomes $\lambda^2 + a_1\lambda + a_2 = 0$, and to be assured of the existence of solutions we need to extend our number system from a real axis to a complex plane. If the roots are indeed complex (and both a_1 and a_2 are real) they will necessarily occur as a complex conjugate pair, as in Example 4.

One might wonder if a further extension of the number system, beyond the complex plane, is required to assure the existence of solutions to (40) for $n \geq 3$. However, it turns out that the complex plane continues to suffice. The characteristic equation (40) necessarily admits n roots. As for the case $n = 2$, they need not be distinct and they need not be real, but if there are complex roots then they necessarily occur in complex conjugate pairs (if all of the a_j's are real).

In this subsection we limit attention to the case where there are n distinct roots of (40), which we denote as $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then each of the exponentials $e^{\lambda_1 x}, \ldots, e^{\lambda_n x}$ is a solution of (1) and, by Theorem 3.3.3,

$$y(x) = C_1 e^{\lambda_1 x} + \cdots + C_n e^{\lambda_n x} \tag{41}$$

is a general solution of (1) if and only if the set of exponentials is LI.

THEOREM 3.4.1 *Linear Independence of a Set of Exponentials*
Let $\lambda_1, \ldots, \lambda_n$ be any numbers, real or complex. The set $\left\{e^{\lambda_1 x}, \ldots, e^{\lambda_n x}\right\}$ is LI (on any given interval I) if and only if the λ's are distinct.

Proof: Recall from Theorem 3.2.2 that if the Wronskian determinant

$$W\left[e^{\lambda_1 x},\dots,e^{\lambda_n x}\right](x)=\begin{vmatrix} e^{\lambda_1 x} & \cdots & e^{\lambda_n x} \\ \lambda_1 e^{\lambda_1 x} & \cdots & \lambda_n e^{\lambda_n x} \\ \vdots & & \vdots \\ \lambda_1^{n-1}e^{\lambda_1 x} & \cdots & \lambda_n^{n-1}e^{\lambda_n x} \end{vmatrix} \tag{42}$$

is not identically zero on I, then the set is LI on I. According to the properties of determinants (Section 10.4), we can factor $e^{\lambda_1 x}$ out of the first column, $e^{\lambda_2 x}$ out of the second, and so on, so that we can re-express W as

$$W\left[e^{\lambda_1 x},\dots,e^{\lambda_n x}\right](x)=e^{(\lambda_1+\cdots+\lambda_n)x}\begin{vmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{vmatrix}. \tag{43}$$

The exponential function on the right-hand side is nonzero. Further, the determinant is of Vandermonde type (Section 10.4), a key property of which is that it is nonzero if the λ's are distinct, as indeed has been assumed here. Thus, W is nonzero (on any interval), so the given set is LI.

Conversely, if the λ's are not distinct, then surely the set is LD because at least two of its members are identical. ■

Consider the following examples.

EXAMPLE 5. The equation

$$y'''-8y'+8y=0 \tag{44}$$

has the characteristic equation $\lambda^3-8\lambda+8=0$. Trial and error reveals that $\lambda=2$ is one root. Hence we can factor $\lambda^3-8\lambda+8$ as $(\lambda-2)p(\lambda)$, where $p(\lambda)$ is a quadratic function of λ. To find $p(\lambda)$ we divide $\lambda-2$ into $\lambda^3-8\lambda+8$, by long division, and obtain $p(\lambda)=\lambda^2+2\lambda-4$ which, in turn, can be factored as $[\lambda-(-1+\sqrt{5})][\lambda-(-1-\sqrt{5})]$. Thus, λ equals 2 and $-1\pm\sqrt{5}$, so

$$\begin{aligned} y(x)&=C_1e^{2x}+C_2e^{(-1+\sqrt{5})x}+C_3e^{(-1-\sqrt{5})x} \\ &=C_1e^{2x}+e^{-x}\left(C_2e^{\sqrt{5}x}+C_2e^{-\sqrt{5}x}\right) \end{aligned} \tag{45}$$

is a general solution of (44).

COMMENT. Alternative to long division, we can find $p(\lambda)$ by writing $\lambda^3-8\lambda+8=(\lambda-2)(a\lambda^2+b\lambda+c)=a\lambda^3+(b-2a)\lambda^2+(c-2b)\lambda-2c$ and determining a,b,c so that coefficients of like powers of λ match on both sides of the equation. ■

EXAMPLE 6. The equation

$$y'''-y=0 \tag{46}$$

has the characteristic equation $\lambda^3 - 1 = 0$, which surely has the root $\lambda = 1$. Thus, $\lambda^3 - 1$ is $\lambda - 1$ times a quadratic function of λ, which function can be found, as above, by long division. Thus, we obtain

$$(\lambda - 1)\left(\lambda^2 + \lambda + 1\right) = 0,$$

so λ equals 1 and $(-1 \pm \sqrt{3}i)/2$. Hence

$$\begin{aligned} y(x) &= C_1 e^x + C_2 e^{(-1+\sqrt{3}i)x/2} + C_3 e^{(-1-\sqrt{3}i)x/2} \\ &= C_1 e^x + e^{-x/2}\left(C_2 e^{i\sqrt{3}x/2} + C_3 e^{-i\sqrt{3}x/2}\right) \\ &= C_1 e^x + e^{-x/2}\left(C_2' \cos \frac{\sqrt{3}}{2}x + C_3' \sin \frac{\sqrt{3}}{2}x\right), \end{aligned} \tag{47}$$

where C_1', C_2', C_3' are arbitrary constants. (Of course, we don't really need the primes in the final answer.) ▮

EXAMPLE 7. The equation

$$y^{(v)} - 7y''' + 12y' = 0 \tag{48}$$

has the characteristic equation $\lambda^5 - 7\lambda^3 + 12\lambda = 0$ or, $\lambda(\lambda^4 - 7\lambda^2 + 12) = 0$. The λ factor gives the root $\lambda = 0$. The quartic factor is actually a quadratic in λ^2, so the quadratic equation gives $\lambda^2 = 4$ and $\lambda^2 = 3$. Thus, λ equals 0, ± 2, $\pm\sqrt{3}$, so

$$y(x) = C_1 + C_2 e^{2x} + C_3 e^{-2x} + C_4 e^{\sqrt{3}x} + C_5 e^{-\sqrt{3}x} \tag{49}$$

is a general solution of (48). ▮

EXAMPLE 8. The equation

$$y^{(iv)} + ky = 0 \tag{50}$$

arises in studying the deflected shape $y(x)$ of a beam on an elastic foundation, where k is a known positive physical constant. Since the characteristic equation $\lambda^4 + k = 0$ gives $\lambda^4 = -k$, to find λ we need to evaluate $(-k)^{1/4}$. The general result is that $z^{1/n}$, for any complex number $z = a + ib$ and any integer $n > 1$, has n values in the complex plane. These values are equally spaced on a circle of radius $r = \sqrt{a^2 + b^2}$ centered at the origin of the complex plane, as is explained in Section 22.4. For our present purpose, let it suffice to merely give the result: $\lambda = (-k)^{1/4} = \pm k^{1/4}\,\dfrac{1+i}{\sqrt{2}}$ and $\pm k^{1/4}\,\dfrac{1-i}{\sqrt{2}}$, so

$$\begin{aligned} y(x) &= C_1 e^{k^{1/4}(1+i)x/\sqrt{2}} + C_2 e^{k^{1/4}(1-i)x/\sqrt{2}} \\ &\quad + C_3 e^{k^{1/4}(-1+i)x/\sqrt{2}} + C_4 e^{k^{1/4}(-1-i)x/\sqrt{2}} \\ &= e^{k^{1/4}x/\sqrt{2}}\left(C_1' \cos \frac{k^{1/4}}{\sqrt{2}}x + C_2' \sin \frac{k^{1/4}}{\sqrt{2}}x\right) \\ &\quad + e^{-k^{1/4}x/\sqrt{2}}\left(C_3' \cos \frac{k^{1/4}}{\sqrt{2}}x + C_4' \sin \frac{k^{1/4}}{\sqrt{2}}x\right) \end{aligned} \tag{51}$$

is a general solution of (50). ▮

3.4.4. Repeated roots. Thus far we have considered only the generic case, where the nth-order characteristic equation (40) admits n distinct roots $\lambda_1, \dots, \lambda_n$. To complete our discussion, we need to consider the case where one or more of the roots is repeated. We say that a root λ_j of (40) is **repeated** if (40) contains the factor $\lambda - \lambda_j$ more than once. More specifically, we say that λ_j is a **root of order k** if (40) contains the factor $\lambda - \lambda_j$ k times. For instance, if the characterisitic equation for some given sixth-order equation can be factored as $(\lambda + 2)(\lambda - 5)^3(\lambda - 1)^2 = 0$, then the roots $\lambda = 5$ and $\lambda = 1$ are repeated; $\lambda = 5$ is a root of order 3 and $\lambda = 1$ is a root of order 2. We can say that

$$y(x) = C_1 e^{-2x} + C_2 e^{5x} + C_3 e^{x}$$

is a solution for any constants C_1, C_2, C_3, but the latter falls short of being a *general* solution of the sixth-order differential equation since it is not a linear combination of six LI solutions. The problem, in such a case of repeated roots, is how to find the missing solutions. Evidently, they will not be of the form $e^{\lambda x}$, for if they were then we would have found them when we sought $y(x)$ in that form.

We will use a simple example to show how to obtain such "missing solutions," and will then state the general result as a theorem.

EXAMPLE 9. *Reduction of Order.* The equation

$$y'' + 2y' + y = 0 \tag{52}$$

has the characteristic equation $\lambda^2 + 2\lambda + 1 = (\lambda + 1)^2 = 0$, so $\lambda = -1$ is a root of order 2. Thus, we have the solution Ae^{-x} but are missing a second linearly independent solution, which is needed if we are to obtain a general solution of (52).

To find the missing solution, we use Lagrange's method of **reduction of order**, which works as follows. Suppose that we know one solution, say $y_1(x)$, of a given linear homogeneous differential equation, and we seek one or more other linearly independent solutions. If $y_1(x)$ is a solution then, of course, so is $Ay_1(x)$, where A is an arbitrary constant. According to the method of reduction of order, we let A vary and seek $y(x)$ in the form $y(x) = A(x)y_1(x)$. Putting that form into the given differential equation results in another differential equation on the unknown $A(x)$, but that equation inevitably will be simpler than the original differential equation on y, as we shall see.

In the present example, $y_1(x)$ is e^{-x}, so to find the missing solution we seek

$$y(x) = A(x)e^{-x}. \tag{53}$$

From (53), $y' = (A' - A)e^{-x}$ and $y'' = (A'' - 2A' + A)e^{-x}$, and putting these expressions into (52) gives

$$(A'' - 2A' + A + 2A' - 2A + A)\, e^{-x} = 0, \tag{54}$$

so that $A(x)$ must satisfy the second-order differential equation obtained by equating the coefficient of e^{-x} in (54) to zero, namely, $A'' - 2A' + A + 2A' - 2A + A = 0$. The cancellation of the three A terms in that equation is not a coincidence, for if $A(x)$ were a constant [in which case the A' and A'' terms in (54) would drop out] then the terms on the left-hand side of (54) would have to cancel to zero because Ae^{-x} is a solution of the original homogeneous differential equation if A is a constant. Thanks to that (inevitable) cancellation, the differential equation governing $A(x)$ will be of the form

$$A'' + \alpha A' = 0, \tag{55}$$

for some constant α, and this second-order equation can be reduced to the first-order equation $v' + \alpha v = 0$ by setting $A' = v$; hence the name reduction of order for the method. In fact, not only do the A terms cancel, as they must, the A' terms happen to cancel as well, so in place of (55) we have the even simpler equation

$$A'' = 0 \tag{56}$$

on $A(x)$. Integration gives $A(x) = C_1 + C_2 x$, so that (53) becomes

$$y(x) = C_1 e^{-x} + C_2 x e^{-x}. \tag{57}$$

The $C_1 e^{-x}$ term merely reproduces that which was already known (recall the second sentence of this example), and the $C_2 x e^{-x}$ term is the desired missing solution. Since the two are LI, (57) is a general solution of (52). ∎

Similarly, suppose we have an eighth-order equation, the characteristic equation of which can be factored as $(\lambda - 2)^3(\lambda + 1)^4(\lambda + 5)$, say, so that 2 is a root of order 3 and -1 is a root of order 4. If we take the solution Ae^{2x} associated with the root $\lambda = 2$, and apply reduction of order by seeking y in the form $A(x)e^{2x}$, then we obtain $A''' = 0$ and $A(x) = C_1 + C_2 x + C_3 x^2$ and hence the "string" of solutions $C_1 e^{2x}, C_2 x e^{2x}, C_3 x^2 e^{2x}$ coming from the root $\lambda = 2$. Likewise, if we take the solution Ae^{-x} associated with the root $\lambda = -1$, and apply reduction of order, we obtain $A(x) = C_4 + C_5 x + C_6 x^2 + C_7 x^3$ and hence the string of solutions $C_4 e^{-x}, C_5 x e^{-x}, C_6 x^2 e^{-x}, C_7 x^3 e^{-x}$ coming from the root $\lambda = -1$, so that we have a general solution

$$\begin{aligned} y(x) &= \left(C_1 + C_2 x + C_3 x^2\right) e^{2x} \\ &\quad + \left(C_4 + C_5 x + C_6 x^2 + C_7 x^3\right) e^{-x} + C_8 e^{-5x} \end{aligned} \tag{58}$$

of the original differential equation. [To verify that this is indeed a general solution one would need to show that the eight solutions contained within (58) are LI, as could be done by working out the Wronskian W and showing that $W \neq 0$.]

EXAMPLE 10. For

$$y'''' - y'' = 0 \tag{59}$$

the characteristic equation $\lambda^4 - \lambda^2 = 0$ gives $\lambda = 0, 0, 1, -1$ and hence the solution $y(x) = A + Be^{x} + Ce^{-x}$. The latter falls short of being a general solution of (59) because the repeated root $\lambda = 0$ gave the single solution A. To find the missing solution by reduction of order we could vary the parameter A and seek $y(x) = A(x)$, but surely there can be no gain in that step since it merely amounts to a name change, from $y(x)$ to $A(x)$. This situation will always occcur when the repeated root is zero, but in that case we can achieve a reduction of order more directly. In the case of (59) we can set $y'' = p$. Then the fourth-order equation (59) is reduced to the second-order equation $p'' - p = 0$, so

$$p(x) = Ae^{x} + Be^{-x}.$$

But $y'' = p$, so

$$y'(x) = \int p(x)\,dx = Ae^{x} - Be^{-x} + C.$$

Hence

$$y(x) = \int \left(Ae^{x} - Be^{-x} + C\right) dx = Ae^{x} + Be^{-x} + Cx + D$$

is the general solution of (59). Observe that the pattern is the same: the repeated root $\lambda = 0$ gives the solution $(C_1 + C_2 x)e^{0x}$, where C_1 is D and C_2 is C. ▮

We organize these results as the following theorem.

THEOREM 3.4.2 *Repeated Roots of Characteristic Equation*
If λ_1 is a root of order k, of the characteristic equation (40), then $e^{\lambda_1 x}, xe^{\lambda_1 x}, \dots, x^{k-1}e^{\lambda_1 x}$ are k LI solutions of the differential equation (1).

Proof: Denote (1) in operator form as $L[y] = 0$, where

$$L = \frac{d^{(n)}}{dx^{(n)}} + a_1 \frac{d^{(n-1)}}{dx^{(n-1)}} + \cdots + a_{n-1}\frac{d}{dx} + a_n. \tag{60}$$

Then

$$L\left[e^{\lambda x}\right] = \left(\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n\right)e^{\lambda x},$$

or

$$L\left[e^{\lambda x}\right] = (\lambda - \lambda_1)^k p(\lambda)e^{\lambda x}, \tag{61}$$

where $p(\lambda)$ is a polynomial in λ, of degree $n - k$. Since (61) holds for all λ, we can set $\lambda = \lambda_1$ in that formula. Doing so, the right-hand side of (61) vanishes, so that $L\left[e^{\lambda_1 x}\right] = 0$ and hence $e^{\lambda_1 x}$ is a solution of $L[y] = 0$. Our object, now, is to show that $xe^{\lambda_1 x}, \dots, x^{k-1}e^{\lambda_1 x}$ are solutions as well.

To proceed, differentiate (61) with respect to λ (λ, not x):

$$\frac{d}{d\lambda}L\left[e^{\lambda x}\right] = k(\lambda - \lambda_1)^{k-1}p(\lambda)e^{\lambda x} + (\lambda - \lambda_1)^k \frac{d}{d\lambda}\left(p(\lambda)e^{\lambda x}\right). \tag{62}$$

The left-hand side of (62) calls for $e^{\lambda x}$ to be differentiated first with respect to x, according to the operator L defined in (60) and then with respect to λ. Since we can interchange the order of these differentiations, we can express the left-hand side as $L\left[\frac{d}{d\lambda}e^{\lambda x}\right]$, that is, as $L\left[xe^{\lambda x}\right]$. Thus, one differentiation of (61) with respect to λ gives

$$L\left[xe^{\lambda x}\right] = k\left(\lambda - \lambda_1\right)^{k-1} p(\lambda)e^{\lambda x} + \left(\lambda - \lambda_1\right)^k \frac{d}{d\lambda}\left(p(\lambda)e^{\lambda x}\right). \tag{63}$$

Setting $\lambda = \lambda_1$ in (63) gives $L\left[xe^{\lambda_1 x}\right] = 0$. Hence, not only is $e^{\lambda_1 x}$ a solution, so is $xe^{\lambda_1 x}$. Repeated differentiation with respect to λ reveals that $x^2 e^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are solutions as well, as was to be proved.

That's as far as we can go because at that point one more differentiation would give a leading term of $k!\left(\lambda - \lambda_1\right)^0 p(\lambda_1)e^{\lambda_1 x}$ plus terms with factors of $\left(\lambda - \lambda_1\right)^1, \left(\lambda - \lambda_1\right)^2, \ldots, \left(\lambda - \lambda_1\right)^k$ on the right-hand side. The latter terms vanish for $\lambda = \lambda_1$, but the leading term does not because $p(\lambda_1) \neq 0$ (because $\lambda - \lambda_1$ is not among the factors of p) and $e^{\lambda_1 x} \neq 0$.

Verification that the solutions $e^{\lambda_1 x}, xe^{\lambda_1 x}, \ldots, x^{k-1}e^{\lambda_1 x}$ are LI is left for the exercises. ■

EXAMPLE 11. As a final example, consider the equation

$$y^{(iv)} - 8y''' + 26y'' - 40y' + 25y = 0 \tag{64}$$

with characteristic equation $\lambda^4 - 8\lambda^3 + 26\lambda^2 - 40\lambda + 25 = 0$ and repeated complex roots $\lambda = 2+i,\ 2+i,\ 2-i,\ 2-i$. It follows that

$$\begin{aligned} y(x) &= \left(C_1 + C_2 x\right) e^{(2+i)x} + \left(C_3 + C_4 x\right) e^{(2-i)x} \\ &= e^{2x}\left[\left(C_1 e^{ix} + C_3 e^{-ix}\right) + x\left(C_2 e^{ix} + C_4 e^{-ix}\right)\right] \\ &= e^{2x}\left[\left(A\cos x + B\sin x\right) + x\left(C\cos x + D\sin x\right)\right] \end{aligned}$$

is a general solution of (64). ■

3.4.5. Stability. An important consideration in applications, especially feedback control systems, is whether or not a system is "stable." Normally, stability has to do with the behavior of a system over time, so let us change the name of the independent variable from x to t in (1):

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1}y}{dt^{n-1}} + \cdots + a_{n-1}\frac{dy}{dt} + a_n y = 0, \tag{65}$$

and let us denote the general solution of (65) as $y(t) = C_1 y_1(t) + \cdots + C_n y_n(t)$. We say that the system described by (65) (be it mechanical, electrical, economic, or whatever) is **stable** if all of its solutions are bounded – that is, if there exists a constant M_j for each solution $y_j(t)$ such that $|y_j(t)| < M_j$ for all $t > 0$. If the system is not stable, then it is **unstable**.

THEOREM 3.4.3 *Stability*
For the system described by (65) to be stable, it is necessary and sufficient that the characteristic equation of (65) have no roots to the right of the imaginary axis in the complex plane and that any roots on the imaginary axis be nonrepeated.

Proof: Let $\lambda = a + ib$ be any nonrepeated root of the characteristic equation; we call *a the real part of λ* and write $\text{Re}\lambda = a$, and *b the imaginary part of λ* and write $\text{Im}\lambda = b$. Such a root will contribute a solution $e^{(a+ib)t} = e^{at}(\cos bt + i \sin bt)$. Since the magnitude (modulus, to be more precise) of a complex number $x + iy$ is defined as $|x + iy| = \sqrt{x^2 + y^2}$, and the magnitude of the product of complex numbers is the product of their magnitudes, we see that $\left|e^{(a+ib)t}\right| = \left|e^{at}(\cos bt + i \sin bt)\right| = \left|e^{at}\right| |\cos bt + i \sin bt| = e^{at}\sqrt{\cos^2 bt + \sin^2 bt} = e^{at}$ so that solution will be bounded if and only if $a \leq 0$, that is, if λ does not lie to the right of the imaginary axis.

Next, let $\lambda = a + ib$ be a repeated root of order k, with $a \neq 0$. Such a root will contribute solutions of the form $t^p e^{(a+ib)t} = t^p e^{at}(\cos bt + i \sin bt)$, for $p = 0, \dots, k - 1$, with magnitude $t^p e^{at}$. Surely the latter grows unboundedly if $a > 0$ because both factors do, but its behavior is less obvious if $a < 0$ since then the t^p factor grows and the e^{at} decays. To see which one "wins," one can rewrite the product as t^p/e^{-at} and then apply l'Hôpital's rule p times. Doing so, one finds that the ratio tends to zero as $t \to \infty$. [Recall that l'Hôpital's rule applies to indeterminate forms of the type $0/0$ or ∞/∞, not $(\infty)(0)$; that is why we first rewrite $t^p e^{at}$ in the form t^p/e^{-at}.] The upshot is that such solutions are bounded if $\lambda = a + ib$ lies in the left half plane ($a < 0$), and unbounded if it lies in the right half plane ($a > 0$). If λ lies *on* the imaginary axis ($a = 0$), then $\left|t^p e^{(a+ib)t}\right| = \left|t^p e^{ibt}\right| = |t^p(\cos bt + i \sin bt)| = t^p$, which grows unboundedly.

Our conclusion is that all solutions are bounded if and only if no roots lie to the right of the imaginary axis and no repeated roots lie on the imaginary axis, as was to be proved. ■

One is often interested in being able to determine whether the system is stable or not without actually evaluating the n roots of the characteristic equation (40). There are theorems that provide information about stability based directly upon the a_j coefficients in (40). One such theorem is stated below. Another, the Routh–Hurwitz criterion, is given in the exercises to Section 10.4.

THEOREM 3.4.4 *Coefficients of Mixed Sign*
If the coefficients in (40) are real and of mixed sign (there is at least one positive and at least one negative), then there is at least one root λ with $\text{Re}\lambda > 0$, so the system is unstable.

These theorems are not as important as they were before the availability of computer software that can determine the roots of (40) numerically and with great ease. For instance, using *Maple*, one can obtain all roots of the equation

$$x^5 + 3x^4 - 2x^3 + x^2 + x + 5 = 0$$

simply by using the **fsolve** command. Enter

$$\text{fsolve}(x\hat{\ }5 + 3 * x\hat{\ }4 - 2 * x\hat{\ }3 + x\hat{\ }2 + x + 5 = 0,\ x,\ \text{complex});$$

and return. This gives the following printout of the five solutions:

$$-3.6339286,\ -.58045036 - .797312491I,\ -.58045036 + .797312491I,$$
$$.89741468 - .780568501I,\ .89741468 + .780568501I$$

In this example, observe that there are, indeed, roots with positive real parts, as predicted by Theorem 3.4.4, so the system is unstable.

For equations of fourth degree or lower, such software works even if one or more of the coefficients are unspecified, in which case the roots are given in terms of those parameters.

Closure. In this section we limited our attention to linear homogeneous differential equations with constant coefficients, a case of great importance in applications. Seeking solutions in exponential form, we found the characteristic equation to be central. According to the fundamental theorem of algebra, such equations always have at least one root, so we are guaranteed of finding at least one exponential solution of the differential equation. If the n roots $\lambda_1, \dots, \lambda_n$ are distinct, then each root λ_j contributes a solution $e^{\lambda_j x}$, and their superposition gives a general solution of (1) in the form

$$y(x) = C_1 e^{\lambda_1 x} + \cdots + C_n e^{\lambda_n x}. \tag{66}$$

If any root λ_j is repeated, say of order k, then it contributes not only the solution $e^{\lambda_j x}$, but the k LI solutions $e^{\lambda_j x}, xe^{\lambda_j x}, \dots, x^{k-1}e^{\lambda_j x}$ to the general solution. Thus, in the generic case of distinct roots, the general solution of (1) is of the form (66); in the nongeneric case of repeated roots, the solution also contains one or more terms which are powers of x times exponentials.

It should be striking how simple is the solution process for linear constant-coefficient homogeneous equations, with the only difficulty being algebraic – the need to find the roots of the characteristic equation. The reason for this simplicity is that most of the work was done simply in deciding to look for solutions in the right place, within the set of exponential functions. Also, observe that although in a fundamental sense the solving of a differential equation in some way involves integration, the methods discussed in this section required no integrations, in contrast to most of the methods of solution of first-order equations in Chapter 1.

In the final (optional) section we introduced the concept of stability, and in Theorem 3.4.3 we related the stability of the physical system to the placement of the roots of the characteristic equation in the complex plane.

Computer software. To obtain a general solution of $y''' - 9y' = 0$ using *Maple*, use the command

$$\text{dsolve}(\{\text{diff}(y(x), x, x, x) - 9 * \text{diff}(y(x), x) = 0\}, y(x));$$

and to solve the ODE subject to the initial conditions $y(0) = 5,\ y'(0) = 2,\ y''(0) = -4$, use the command

$$\text{dsolve}(\{\text{diff}(y(x), x, x, x) - 9 * \text{diff}(y(x), x) = 0,\ y(0) = 5,\\ D(y)(0) = 2,\ D(D(y))(0) = -4\}, y(x));$$

In place of diff$(y(x), x, x, x)$ we could use diff$(y(x), x\$3)$, for brevity.

EXERCISES 3.4

1. Use whichever of equations (5)–(8) are needed to derive these relations between the circular and hyperbolic functions:

(a) $\cos(ix) = \cosh x$ (b) $\sin(ix) = i \sinh x$
(c) $\cosh(ix) = \cos x$ (d) $\sinh(ix) = i \sin x$

2. Use equations (6) and/or (7) to derive or verify

(a) equation (9) (b) equation (10)
(c) equation (13a) (d) equation (13b)
(e) equation (13c) (f) equation (13d)

3. Theorem 3.4.2 states that $e^{\lambda_1 x}, xe^{\lambda_1 x}, \dots, x^{k-1}e^{\lambda_1 x}$ are LI. Prove that claim.

4. (*Nonrepeated roots*) Find a general solution of each of the following equations, and a particular solution satisfying the given conditions, if such conditions are given.

(a) $y'' + 5y' = 0$
(b) $y'' - y' = 0$
(c) $y'' + y' = 0;\quad y(0) = 3,\quad y'(0) = 0$
(d) $y'' - 3y' + 2y = 0;\quad y(1) = 1,\quad y'(1) = 0$
(e) $y'' - 4y' - 5y = 0;\quad y(1) = 1, y'(1) = 0$
(f) $y'' + y' - 12y = 0;\quad y(-1) = 2,\quad y'(-1) = 5$
(g) $y'' - 4y' + 5y = 0;\quad y(0) = 2,\quad y'(0) = 5$
(h) $y'' - 2y' + 3y = 0;\quad y(0) = 4,\quad y'(0) = -1$
(i) $y'' - 2y' + 2y = 0;\quad y(0) = 0,\quad y'(0) = -5$
(j) $y'' + 2y' + 3y = 0;\quad y(0) = 0,\quad y'(0) = 3$
(k) $y''' + 3y' - 4y = 0;\quad y(0) = 0,\quad y'(0) = 0,\quad y''(0) = 6$
(l) $y''' - y'' + 2y' = 0;\quad y(0) = 1,\quad y'(0) = 0,\quad y''(0) = 0$
(m) $y''' + y'' - 2y = 0$
(n) $y^{(iv)} - y = 0$
(o) $y^{(iv)} - 2y'' - 3y = 0$
(p) $y^{(iv)} + 6y'' + 8y = 0$
(q) $y^{(iv)} + 7y'' + 12y = 0$
(r) $y^{(iv)} - 2y''' - y'' + 2y' = 0$

5. (a)–(r) Solve the corresponding problem in Exercise 4 using computer software.

6. (*Repeated roots*) Find a general solution of each of the following equations, and a particular solution satisfying the given conditions, if such conditions are given.

(a) $y'' = 0;\quad y(-3) = 5,\quad y'(-3) = -1$
(b) $y'' + 6y' + 9y = 0;\quad y(1) = e,\quad y'(1) = -2$
(c) $y''' = 0;\quad y(0) = 3,\quad y'(0) = -5,\quad y''(0) = 1$
(d) $y''' + 5y'' = 0;\quad y(0) = 1,\quad y'(0) = 0,\quad y''(0) = 0$
(e) $y''' + 3y'' + 3y' + y = 0$
(f) $y''' - 3y'' + 3y' - y = 0$
(g) $y''' - y'' - y' + y = 0$
(h) $y^{(iv)} + 3y''' = 0$
(i) $y^{(iv)} + y''' + y'' = 0$
(j) $y^{(iv)} + 8y'' + 16y = 0$
(k) $y^{(vi)} = 0;\quad y(0) = y'(0) = y''(0) = y'''(0) = y^{(iv)}(0) = 0,\quad y^{(v)}(0) = 3$

7. (a)–(k) Solve the corresponding problem in Exercise 6 using computer software.

8. If the roots of the characteristic equation are as follows, then find the original differential equation and also a general solution of it:

(a) $2, 6$ (b) $2i, -2i$
(c) $4 - 2i, 4 + 2i$ (d) $-2, 3, 5$

(e) $2, 3, -1$ (f) $1, 1, -2$
(g) $4, 4, 4, i, -i$ (h) $1, -1, 2+i, 2-i$
(i) $0, 0, 0, 0, 7, 9$ (j) $1+i, 1+i, 1-i, 1-i$

9. (*Complex a_j's*) Find a general solution of each of the following equations. NOTE: Normally, the a_j coefficients in (1) are real, but the results of this section hold even if they are not (except for Theorem 3.4.4, which explicitly requires that the coefficients be real). However, be aware that if the a_j coefficients are not all real, then complex roots do not necessarily occur in complex conjugate pairs. For instance, $\lambda^2 + 2i\lambda + 1 = 0$ has the roots $\lambda = (\sqrt{2} - 1)i, -(\sqrt{2} + 1)i$.

(a) $y'' - 2iy' + y = 0$ (b) $y'' - 3iy' - 2y = 0$
(c) $y'' + iy' - y = 0$ (d) $y'' - 2iy' - y = 0$
(e) $y'' - iy = 0$ HINT: Verify, and use, the fact that $\sqrt{i} = \pm(1+i)/\sqrt{2}$.
(f) $y''' + 4iy'' - y' = 0$
(g) $y''' + iy' = 0$ HINT: Verify, and use, the fact that $\sqrt{-i} = \pm(1-i)/\sqrt{2}$
(h) $y''' - (1+2i)y'' + (1+i)y' - 2(1+i)y = 0$ HINT: One root is found, by inspection, to be $\lambda = -i$.

10. (a)–(h) Solve the corresponding problem in Exercise 9 using computer software.

11. (*Solution by factorization of the operator*) We motivated the idea of seeking solutions to (1) in the form $e^{\lambda x}$ by observing that the general solution of the first-order equation $y' + a_1 y = 0$ is an exponential, $Ce^{-a_1 x}$, and wondering if higher-order equations might admit exponential solutions too. A more compelling approach is as follows. Having already seen that the first-order equation admits an exponential solution, consider the second-order equation (18).

(a) Show that (18) can be written, equivalently, as

$$(D - \lambda_1)(D - \lambda_2)y = 0, \tag{11.1}$$

where D denotes d/dx, and λ_1 and λ_2 are the two roots of $\lambda^2 + a_1\lambda + a_2 = 0$. NOTE: In (11.1) we accomplish a **factorization** of the original differential operator $L = D^2 + a_1 D + a_2$ as $(D - \lambda_1)(D - \lambda_2)$. By the left-hand side of (11.1), we mean $(D - \lambda_1)((D - \lambda_2)y)$. That is, first let the operator to the left of y (namely, $D - \lambda_2$) act on y, then let the operator to the left of $(D - \lambda_2)y$ (namely, $D - \lambda_1$) act on that.
(b) To solve (11.1), let $(D - \lambda_2)y = u$, so that (18) reduces to the first-order equation

$$\frac{du}{dx} - \lambda_1 u = 0. \tag{11.2}$$

Solve (11.2) for u, put that u on the right-hand side of $\dfrac{dy}{dx} - \lambda_2 y = u$, which is again of first order, and solve the latter for y. Show that if λ_1, λ_2 are distinct, then the result is given by (23), whereas if they are repeated, then the result is $y(x) = (C_1 + C_2 x)e^{\lambda_1 x}$.
(c) Solve $y'' - 3y' + 2y = 0$ by factoring the operator as $(D-1)(D-2)y = 0$. Solve the latter by the method outlined in (b): Setting $(D-2)y \equiv u$, solve $(D-1)u = u' - u = 0$ for $u(x)$. Then, knowing $u(x)$, solve $(D-2)y = u$, namely, $y' - 2y = u(x)$, for $y(x)$.
(d) Same as (c), for $y'' - 4y = 0$.
(e) Same as (c), for $y'' + 4y' + 3y = 0$.
(f) Same as (c), for $y'' + 2y' + y = 0$.
(g) Same as (c), for $y'' + 4y' + 4y = 0$.
NOTE: Similarly for higher-order equations. For instance, $y''' - 2y'' - y' + 2y = (D-2)(D+1)(D-1)y = 0$ can be solved by setting $(D+1)(D-1)y \equiv u$ and solving $(D-2)u = 0$ for $u(x)$; then set $(D-1)y \equiv v$ and solve $(D+1)v = u$ for $v(x)$; finally, solve $(D-1)y = v$ for $y(x)$. The upshot is that the solution of an nth-order linear homogeneous differential equation with constant coefficients can be reduced to the solution of a sequence of n first-order linear equations.

12. Use computer software to obtain the roots of the given characteristic equation, state whether the system is stable or unstable, and explain why. If Theorem 3.4.4 applies, then show that your results are consistent with the predictions of that theorem.

(a) $\lambda^3 - 3\lambda^2 + 26\lambda - 2 = 0$
(b) $\lambda^3 + 3\lambda^2 + 2\lambda + 2 = 0$
(c) $\lambda^4 + \lambda^3 + 3\lambda^2 + 2\lambda + 2 = 0$
(d) $\lambda^4 + \lambda^3 + 5\lambda^2 + \lambda + 4 = 0$
(e) $\lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 - \lambda^2 + \lambda + 3 = 0$
(f) $\lambda^6 + 9\lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(g) $\lambda^6 + \lambda^5 + 5\lambda^4 + 4\lambda^3 + 4\lambda^2 + 8\lambda + 4 = 0$
(h) $\lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(i) $\lambda^8 - \lambda^6 + \lambda^5 + 5\lambda^4 + 2\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$
(j) $\lambda^8 + \lambda^7 + \lambda^6 + \lambda^5 + 5\lambda^4 + 21\lambda^3 + 7\lambda^2 + \lambda + 3 = 0$

3.5 Application to Harmonic Oscillator: Free Oscillation

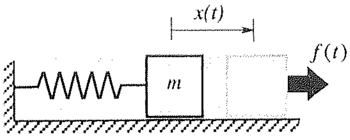


Figure 1. Mechanical oscillator.

In Section 1.3 we discussed the modeling of the mechanical oscillator reproduced here in Fig. 1. Neglecting air resistance, the block of mass m is subjected to a restoring force due to the spring, a "drag" force due to the friction between the block and the lubricated table top, and an applied force $f(t)$. (By a restoring force, we mean that the force opposes the stretch or compression in the spring.) Most of that discussion focused on the modeling of the spring force and friction force, and we derived the approximate equation of motion

$$mx'' + cx' + kx = f(t), \tag{1}$$

where c is the *damping coefficient*, k is the *spring stiffness*. Besides the differential equation, let us regard the initial displacement and initial velocity,

$$x(0) = x_0 \quad \text{and} \quad x'(0) = x_0', \tag{2}$$

respectively, as specified values.

In this section we consider the solution for the case where $f(t) = 0$:

$$mx'' + cx' + kx = 0. \tag{3}$$

This is the so-called *unforced*, or *free*, oscillation. According to Theorem 3.3.1, the solution $x(t)$ to (3) and (2) does exist and is unique. To find it, we seek $x(t) = e^{\lambda t}$ and obtain the characteristic equation $m\lambda^2 + c\lambda + k = 0$, with roots

$$\lambda = \frac{-c \pm \sqrt{c^2 - 4mk}}{2m}. \tag{4}$$

Consider first the case where there is no damping, so $c = 0$ and (3) becomes

$$\boxed{mx'' + kx = 0.} \tag{5}$$

That is, the friction is small enough so that it can be neglected altogether. Then (4) gives $\lambda = \pm i\sqrt{k/m}$, and the solution of (5) is

$$\boxed{x(t) = Ae^{i\omega t} + Be^{-i\omega t},} \tag{6}$$

where $\omega = \sqrt{k/m}$ is the so-called **natural frequency** of the system, in rad/sec. Or, equivalent to (6) and favored in this text,

$$\boxed{x(t) = C\cos\omega t + D\sin\omega t.} \tag{7}$$

In fact, there is another useful form of the general solution, namely,

$$\boxed{x(t) = E\sin(\omega t + \phi),} \tag{8}$$

where the integration constants E and ϕ can be determined in terms of C and D as follows. To establish the equivalence of (8) and (7), recall the trigonometric identity $\sin{(A+B)} = \sin B \cos A + \sin A \cos B$. Then

$$E \sin{(\omega t + \phi)} = E \sin\phi \cos\omega t + E \cos\phi \sin\omega t,$$

which is identical to $C \cos\omega t + D \sin\omega t$ if

$$C = E \sin\phi \quad \text{and} \quad D = E \cos\phi. \tag{9a,b}$$

Squaring and adding equations (9), and also dividing one by the other, gives

$$\boxed{E = \sqrt{C^2 + D^2} \quad \text{and} \quad \phi = \tan^{-1}\frac{C}{D},} \tag{10a,b}$$

respectively, as the connection between the equivalent forms (7) and (8). It will be important to be completely comfortable with the equivalence between the solution forms (6), (7), and (8). Both the square root and the $\tan^{-1}$ in (10) are multi-valued. We will understand the square root as the positive one and the $\tan^{-1}$ to lie between $-\pi$ and π. Specifically, it follows from (9), with $E > 0$, that if $C > 0$ and $D > 0$ then $0 < \phi < \pi/2$, if $C > 0$ and $D < 0$ then $\pi/2 < \phi < \pi$, if $C < 0$ and $D > 0$ then $-\pi/2 < \phi < 0$, and if $C < 0$ and $D < 0$ then $-\pi < \phi < -\pi/2$.

For instance, consider $6\cos t - 2\sin t$. Then $E = \sqrt{36+4} = \sqrt{40}$ and $\phi = \tan^{-1}\left(\frac{+6}{-2}\right)$. A calculator or typical computer software will interpret $\tan^{-1}(\;)$ as $-\pi/2 < \tan^{-1}(\;) < \pi/2$, namely, in the first or fourth quadrant. Not able to distinguish $(+6)/(-2)$ from $(-6)/(+2)$, it will give $\tan^{-1}\left(-\frac{6}{2}\right) = -1.25$ rad, which is incorrect. The correct value is in the second quadrant, namely, $\phi = \pi - 1.25 = 1.89$ rad. Thus, $6\cos t - 2\sin t = \sqrt{40}\sin{(t + 1.89)}$.

Whereas C and D in (7) have no special physical significance, E and ϕ in (8) are the **amplitude** and **phase angle** of the vibration, respectively (Fig. 2a).

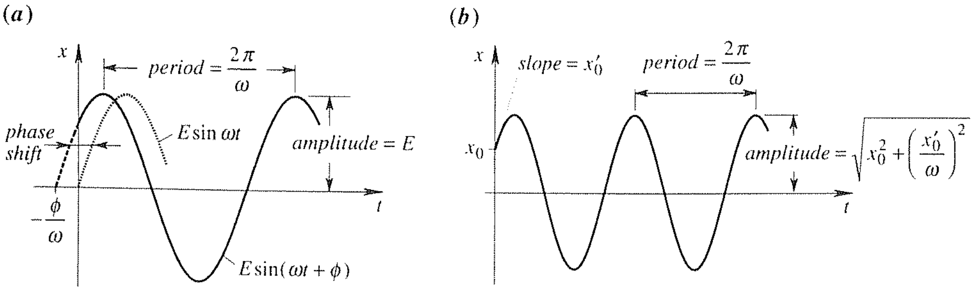


Figure 2. (a) Graphical significance of ω, ϕ. (b) Undamped free oscillation.

Although (8) is advantageous conceptually, in that the amplitude E and phase angle ϕ are physically and graphically meaningful, it is a bit easier to apply the initial conditions to (7):

$$x(0) = x_0 = C, \quad x'(0) = x_0' = \omega D$$

so $C = x_0, D = x_0'/\omega$, and the solution is

$$x(t) = x_0 \cos \omega t + \frac{x_0'}{\omega} \sin \omega t, \tag{11}$$

a plot of which is shown in Fig. 2b for representative initial conditions x_0 and x_0'.

Before continuing, consider the relationship between the mathematics and the physics. For example, the frequency $\omega = \sqrt{k/m}$ increases with k, decreases with m, and is independent of the initial conditions x_0 and x_0', and hence the amplitude which, according to (11) and (10), is $\sqrt{x_0^2 + (x_0'/\omega)^2}$. Do these results make sense? Probably the increase of ω with k fits with our experience with springs, and its decrease with m makes sense as well. However, one might well expect the frequency to vary with the amplitude of the vibration. We will come back to this point in Chapter 7, where we consider more realistic *non*linear models.

Now suppose there is some damping, $c > 0$. From (4) we see that there are three cases of interest. If we define the **critical damping** as $c_{cr} = \sqrt{4mk} = 2\sqrt{mk}$, then the solution is qualitatively different depending upon whether $c < c_{cr}$ (the "underdamped" case), $c = c_{cr}$ (the "critically damped" case), or $c > c_{cr}$ (the "overdamped" case).

Underdamped vibration ($c < c_{cr}$). In this case (4) gives two complex conjugate roots

$$\lambda = \frac{1}{2m}\left(-c \pm \sqrt{c^2 - c_{cr}^2}\right) = \frac{1}{2m}\left(-c \pm i\sqrt{c_{cr}^2 - c^2}\right)$$
$$= -\frac{c}{2m} \pm i\sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}$$

so a general solution of (1) is

$$x(t) = e^{-\frac{c}{2m}t}\left[A \cos \sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\, t + B \sin \sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\, t\right], \tag{12}$$

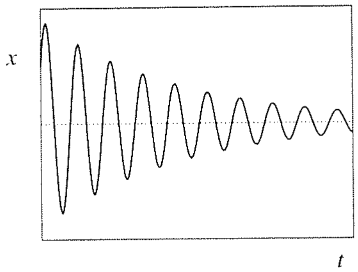


Figure 3. Underdamped free oscillation.

where A and B can be determined from the initial conditions (2). Of course, we could express the bracketed part in the form (8) if we like.

Comparing (7) and (12), observe that the damping has two effects. First, it introduces the $e^{-(c/2m)t}$ factor, which causes the oscillation to "damp out" as $t \to \infty$, as illustrated in Fig. 3. That is, the amplitude tends to zero as $t \to \infty$. Second, it reduces the frequency from the natural frequency ω to $\sqrt{\omega^2 - (c/2m)^2}$; that is, it makes the system more sluggish, as seems intuitively reasonable. (It might appear from Fig. 2b and 3 that the damping *in*creases the frequency, but that appearance is only because we have compressed the t scale in Fig. 3.)

Critically damped vibration ($c = c_{cr}$). As c is increased further, the system becomes so sluggish that when c attains the critical value c_{cr} the oscillation ceases altogether. In this case (4) gives the repeated root $\lambda = -c/2m$, of order two, so

$$x(t) = (A + Bt)e^{-\frac{c}{2m}t}. \tag{13}$$

Although the t in $A+Bt$ grows unboundedly, the exponential function decays more powerfully (as discussed within the proof of Theorem 3.4.3) and the solution (13) decays without oscillation, as shown in the $c = c_{cr}$ part of Fig. 4.

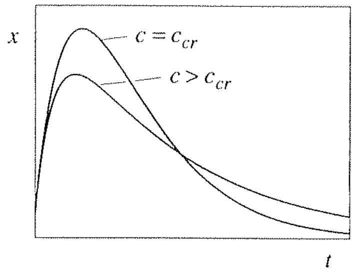


Figure 4. Critically damped and overdamped cases.

Overdamped vibration ($c > c_{cr}$). As c increases beyond c_{cr}, (4) once again gives two distinct roots, but now they are both real and negative (because the $\sqrt{c^2 - 4mk}$ is smaller than c), so

$$x(t) = e^{-\frac{c}{2m}t}\left[A\cosh\sqrt{\left(\frac{c}{2m}\right)^2 - \omega^2}\; t + B\sinh\sqrt{\left(\frac{c}{2m}\right)^2 - \omega^2}\; t\right], \tag{14}$$

where A and B can be determined from the initial conditions (2). Indeed, if one or both roots were positive then we could have exponential growth, which would make no sense, physically. If that did happen we should expect that either there is an error in our mathematics or that our mathematical modeling of the phenomenon is grossly inaccurate.

A representative plot of that solution is shown in the $c > c_{cr}$ part of Fig. 4. For the sake of comparison we have used the same initial conditions to generate the three plots in Figures 3 and 4. Though one can use positive and negative exponentials within the parentheses in (14), in place of the hyperbolic cosine and sine, the latter are more convenient for the application of the initial conditions since the sinh is zero at $t = 0$ and so is the derivative of the cosh.

This completes our solution of equation (3), governing the free oscillation of the mechanical oscillator shown in Fig. 1. It should be emphasized that Fig. 1 is intended only as a schematic equivalent of the actual physical system. For instance, suppose the actual system consists of a beam cantilevered downward, with a mass m at its end, as shown in Fig. 5a. We assume the mass of the beam to be negligible compared to m. It is known from Euler beam theory that if we apply a constant force F, as shown in Fig. 5b, then the end deflection x is given by $x = FL^3/(3EI)$, where L is the length of the beam and EI is its "stiffness" (E is Young's modulus of the material and I is a cross-sectional moment of inertia). Re-expressing the latter as $F = (3EI/L^3)x$, we see that it is of the form $F = kx$, as for a linear spring of stiffness k. Thus, insofar as the modeling and analysis is concerned, the physical beam system is equivalent to the mass-spring arrangement shown in Fig. 6c, where $k_{eq} = 3EI/L^3$ is the stiffness of the equivalent spring and where there is no friction between the block and the table top. The governing equation of motion is

$$mx'' + k_{eq}x = 0. \tag{15}$$

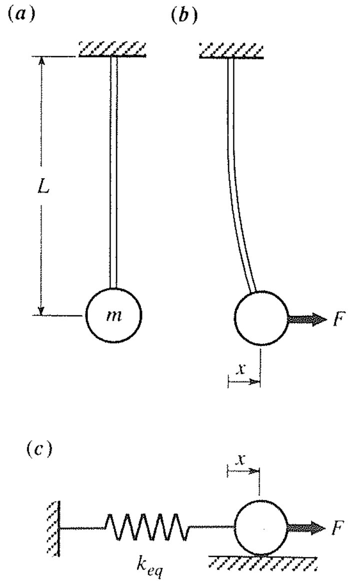


Figure 5. Equivalent mechanical systems.

Just as we neglected the mass of the beam, compared to m, likewise let us neglect the mass of the spring compared to m. (How to account for that mass, approximately, is discussed in the exercises.) It should be noted that, in addition, we are neglecting the rotational motion of the mass, in Fig. 5b, since we have already limited ourselves to the case of small deflections of the beam.

Finally, it has already been pointed out, in Section 2.3, that the force-driven mechanical oscillator is analogous to the voltage-driven RLC electrical circuit reproduced here in Fig. 6, under the equivalence

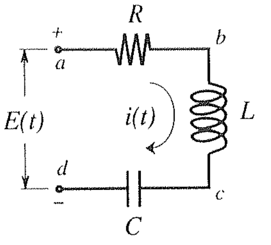


Figure 6. Electrical oscillator; RLC circuit.

$$L \leftrightarrow m,\ R \leftrightarrow c,\ \frac{1}{C} \leftrightarrow k,\ i(t) \leftrightarrow x(t),\ \frac{dE(t)}{dt} \leftrightarrow F(t), \tag{16}$$

so whatever results we have obtained in this section for the mechanical oscillator apply equally well to the electrical oscillator shown in Fig. 6, according to the equivalence given above.

Closure. In this section we have considered the free oscillations of the mechanical harmonic oscillator. We found that for the undamped case the solution is a pure sine wave with an amplitude and phase shift that depend upon the initial conditions – that is, the solution is "harmonic." In the presence of light damping (i.e., for $c < c_{cr}$), the solution suffers exponential decay and a reduction in frequency, these effects becoming more pronounced as c is increased. When c reaches a critical value c_{cr} the oscillation ceases altogether, and as c is increased further the exponential decay is increasingly pronounced.

It should be emphasized that by the damped harmonic oscillator we mean a system that can be modeled by a linear equation of the form $mx'' + cx' + kx = 0$. In most applications, however, the restoring force can be regarded as a linear function of x (namely, kx) only for motions that are sufficiently small departures from an equilibrium configuration; if the motion is not sufficiently small, then one must deal with a more difficult nonlinear differential equation. Thus, for the harmonic oscillator, damped or not, we are able to generate simple closed form solutions, as we have done in this section. For nonlinear oscillators one often gives up on the possibility of finding closed form analytical solutions and relies instead on numerical simulation, as will be discussed in Chapter 6. To illustrate how such nonlinear oscillators arise in applications, we have included several such examples in the exercises that follow.

In terms of formulas, the equivalence of the three forms (6), (7), and (8) should be clearly understood and remembered. In a given application we will use whichever of these seems most convenient, usually (7).

EXERCISES 3.5

1. Re-express each expression in the form $E \sin(\omega t + \phi)$; that is, evaluate E, ϕ, ω.

(a) $6 \cos t + \sin t$

(b) $3 \cos 6t - 4 \sin 6t$

(c) $5 \cos 2t - 12 \sin 2t$

(d) $-2 \cos 3t + 2 \sin 3t$

(e) $\cos 5t - \sin 5t$

(f) $x_0 \cos \omega t + \dfrac{x_0'}{\omega} \sin \omega t$, from (11)

2. We emphasized the equivalence of the solution forms (6), (7), and (8), and discussed the equations (10a,b) that relate C and D in (7) to E and ϕ in (8). Of course, we could have used the cosine in place of the sine, and expressed

$$x(t) = G \cos(\omega t + \psi) \tag{2.1}$$

instead. Derive formulas analogous to (10a,b), expressing G and ψ in terms of C and D.

3. Apply the initial conditions (2) to the general solution (12), and solve for the integration constants A and B in terms of m, c, ω, x_0 and x_0'.

4. Apply the initial conditions (2) to the general solution (14), and solve for the integration constants A and B in terms of m, c, ω, x_0 and x_0'.

5. Consider an undamped harmonic oscillator governed by the equation $mx'' + kx = 0$, with initial conditions $x(0) = x_0, x'(0) = x_0'$. One might expect the frequency of oscillation to depend on the initial displacement x_0. Does it? Explain.

6. We mentioned in the text that the oscillation ceases altogether when c is increased to c_{cr} or beyond. Let us make that statement more precise: for $c \geq c_{cr}$ the graph of $x(t)$ has at most one "flat spot" (on $0 \leq t < \infty$), that is, where $x' = 0$.

(a) Prove that claim.
(b) Make up a case (i.e., give numerical values of m, c, k, x_0, x_0') where there is *no* flat spot on $0 \leq t < \infty$.
(c) Make up a case where there is one flat spot on $0 \leq t < \infty$.

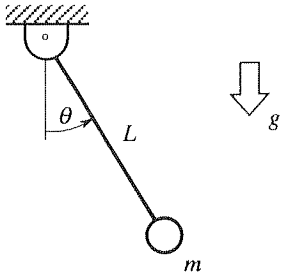
7. (*Logarithmic decrement*) For the underdamped case ($c < c_{cr}$), let x_n and x_{n+1} denote any two successive maxima of $x(t)$.

(a) Show that the ratio $r_n = x_n/x_{n+1}$ is a constant, say r; that is, $x_1/x_2 = x_2/x_3 = \cdots = r$.
(b) Further, show that the natural logarithm of r, called the **logarithmic decrement** δ, is given by $\delta = \dfrac{c}{m} \dfrac{\pi}{\sqrt{\omega^2 - (c/2m)^2}}$.

8. (*Grandfather clock*) Consider a pendulum governed by the equation of motion $mL\theta'' + mg \sin\theta = 0$, or

$$\theta'' + \frac{g}{L} \sin\theta = 0, \tag{8.1}$$

where g is the acceleration of gravity. (See the figure.) If



$|\theta| \ll 1$ (where $\ll$ means much smaller than), then $\sin\theta \approx \theta$, and the nonlinear equation of motion (8.1) can be simplified to the linear equation

$$\theta'' + \frac{g}{L}\theta = 0, \tag{8.2}$$

or, if we allow for some inevitable amount of damping due to friction and air resistance,

$$\theta'' + \epsilon\theta' + \frac{g}{L}\theta = 0, \tag{8.3}$$

where $0 < \epsilon \ll 1$. Now imagine the pendulum to be part of a grandfather's clock. If a ratchet mechanism converts each oscillation to one second of recorded time, how does the clock maintain its accuracy even as it runs down, that is, even when its amplitude of oscillation has diminished to a small fraction of its initial value? Explain.

9. (*Correction for the mass of the spring*) Recall that our model of the mechanical oscillator neglects the effect of the mass of the spring on the grounds that it is sufficiently small compared to that of the mass m. In this exercise we seek to improve our model so as to account, if only approximately, for the mass of the spring. In doing so, we consider the undamped case, for which the equation of motion is $mx'' + kx = 0$.

(a) Multiplying that equation by dx and integrating, derive the "first integral"

$$\frac{1}{2}mx'^2 + \frac{1}{2}kx^2 = C, \tag{9.1}$$

which states that the total energy, the kinetic energy of the mass plus the potential energy of the spring, is a constant.
(b) Let the mass of the spring be m_s. Suppose that the velocity of the elements within the spring at any time t varies linearly from 0 at the fixed end to $x'(t)$ at its attachment to the mass m. Show, subject to that assumption, that the kinetic energy in the spring is $\frac{1}{6}m_s x'^2(t)$. Improving (9.1) to the form

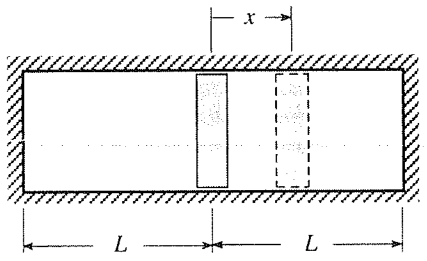
$$\frac{1}{2}\left(m + \frac{1}{3}m_s\right)x'^2 + \frac{1}{2}kx^2 = C, \tag{9.2}$$

obtain, by differentiation with respect to t, the improved equation of motion

$$\left(m+\frac{1}{3}m_s\right)x''+kx=0. \tag{9.3}$$

Thus, as a correction, to take into account the mass of the spring, we merely replace the mass m in $mx''+kx=0$ by an "effective mass" $m+\frac{1}{3}m_s$, which incorporates the effect of the mass of the spring. NOTE: This analysis is approximate in that it assumes the velocity distribution within the spring, whereas that distribution itself needs to be determined, which determination involves the solution of a partial differential equation of wave type, as studied in a later chapter.
(c) In obtaining an effective mass of the form $m+\alpha m_s$, why is it reasonable that α turns out to be less than 1?

10. (*Piston oscillator*) Let a piston of mass m be place at the midpoint of a closed cylinder of cross-sectional area A and length $2L$, as sketched. Assume that the pressure p on either side of the piston satisfies Boyle's law (namely, that the pressure times the volume is constant), and let p_0 be the pressure on both sides when $x=0$.



(a) If the piston is disturbed from its equilibrium position $x=0$, show that the governing equation of motion is

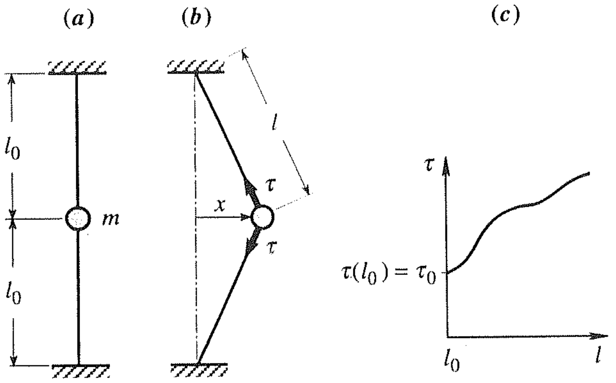
$$mx''+2p_0AL\frac{x}{L^2-x^2}=0. \tag{10.1}$$

(b) Is (10.1) linear or nonlinear? Explain.
(c) Expand the $x/(L^2-x^2)$ term in a Taylor series about $x=0$, up to the third-order term. Keeping only the leading term, derive the linearized version

$$mx''+\frac{2p_0A}{L}x=0 \tag{10.2}$$

of (10.1), which is restricted to the case of small oscillations – that is, where the amplitude of oscillation is small compared to L.
(d) From (10.2), determine the frequency of oscillation, in cycles per second.
(e) Is the resulting linearized model equivalent to the vibration of a mass/spring system, with an equivalent spring stiffness of $k_{eq}=2p_0A/L$? Explain.

11. (*Lateral vibration of a bead on a string*) Consider a mass m, such as a bead, restrained by strings (of negligible mass), in each of which there is a tension τ_0, as shown in Fig. ***a***.



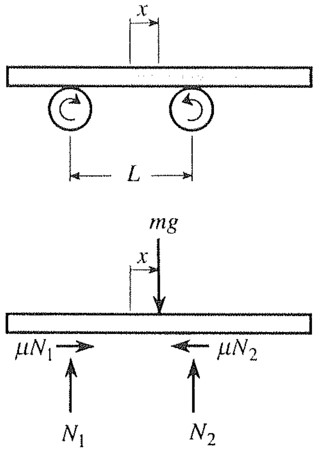
We seek the frequency of small lateral oscillations of m. A lateral displacement x (Fig. ***b***) will cause the length of each string to increase from l_0 to $l(x)=\sqrt{l_0^2+x^2}$. Suppose that the tension τ is found, empirically, to increase with l, from its initial value τ_0, as shown in Fig. ***c***.

(a) Show that the governing equation of lateral motion is

$$mx''+2\frac{\tau\left(\sqrt{l_0^2+x^2}\right)}{\sqrt{l_0^2+x^2}}x=0, \tag{11.1}$$

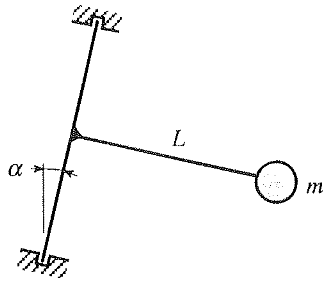
where $\tau\left(\sqrt{l_0^2+x^2}\right)$ is a function, not a product.
(b) Is (11.1) linear or nonlinear? Explain.
(c) Expand the $\tau\left(\sqrt{l_0^2+x^2}\right)x/\sqrt{l_0^2+x^2}$ term in a Taylor series about $x=0$, up to the third-order term. [You should find that the coefficients of these terms involve l_0,τ_0, and $\tau'(l_0)$.]
(d) Linearize the equation of motion by retaining only the leading term of that Taylor series, show that the equivalent spring stiffness is $k_{eq}=2\tau_0/l_0$, and that the frequency of small oscillations is $\frac{1}{2\pi}\sqrt{\frac{2\tau_0}{ml_0}}$ cycles/sec.

12. (*Oscillating platform*) A uniform horizontal platform of mass m is supported by counter-rotating cylinders a distance L apart (see figure). The friction force f exerted on the



platform by each cylinder is proportional to the normal force N between the platform and the cylinder, with constant of proportionality (coefficient of sliding friction) μ: $f = \mu N$. Show that if the cylinder is disturbed from its equilibrium position ($x = 0$), then it will undergo a lateral oscillation of frequency $\omega = \sqrt{2\mu g/L}$ rad/sec, where g is the acceleration of gravity. HINT: Derive the equation of motion governing the lateral displacement x of the midpoint of the platform relative to a point midway between the cylinders.

13. (*Tilted pendulum*) Consider a rod of length L with a point mass m at its end, where the mass of the rod is negligible compared to m. The rod is welded at a right angle to another, which rotates without friction about an axis that is tilted by an angle of α with respect to the vertical (see figure). Let θ denote the angle of rotation of the pendulum, with respect to its equilibrium position (where m is at its lowest possible point, namely, in the plane of the paper).



(a) Derive the governing equation of motion

$$\theta'' + \frac{g}{L}\sin\alpha\sin\theta = 0. \tag{13.1}$$

As a partial check of this result, observe that for $\alpha = \pi/2$ (14.1) does reduce to the equation of motion of the ordinary pendulum (see Exercise 8). HINT: Write down an equation of conservation of energy (kinetic plus potential energy equal a constant), and differentiate it with respect to the time t.

(b) What is the frequency of small amplitude oscillations, in rad/sec? In cycles/sec?

3.6 Solution of Homogeneous Equation: Nonconstant Coefficients

We return to the nth-order linear homogeneous equation

$$a_0(x)\frac{d^ny}{dx^n} + a_1(x)\frac{d^{n-1}y}{dx^{n-1}} + \cdots + a_{n-1}(x)\frac{dy}{dx} + a_n(x)y = 0, \tag{1}$$

and this time allow the a_j coefficients to be nonconstant. The theory of the homogeneous equation, given in Section 3.3, holds whether the coefficients are constant or not. However, the task of finding solutions is generally much more difficult if the coefficients in (1) are not all constants. Only in special cases are we able to find solutions in terms of the elementary functions and in closed form (as opposed, say, to infinite series). This section is devoted to those special cases – most notably, to equations of Cauchy–Euler type. In other cases we generally give up on finding closed-form solutions and, instead, seek solutions in the form of infinite series

(Chapter 4) or pursue a numerical approach (Chapter 6).

3.6.1. Cauchy–Euler equation. If (1) is of the special form

$$\boxed{x^n \frac{d^n y}{dx^n} + c_1 x^{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + c_{n-1} x \frac{dy}{dx} + c_n y = 0,} \tag{2}$$

where the c_j's are constants, it is called a **Cauchy–Euler equation**, and is also called an **equidimensional equation**.

Of most importance to us will be the case where $n = 2$, so let us consider that case first, namely,

$$x^2 y'' + c_1 x y' + c_2 y = 0, \tag{3}$$

and let us consider the x interval to be $0 < x < \infty$; the case of negative x's will be treated separately, below. Suppose we try to solve (3) by seeking y in the form $y = e^{\lambda x}$, where λ is a yet-to-be-determined constant, which form proved successful for the constant-coefficient case. Then $y' = \lambda e^{\lambda x}$ and $y'' = \lambda^2 e^{\lambda x}$, so (3) becomes

$$\lambda^2 x^2 e^{\lambda x} + \lambda c_1 x e^{\lambda x} + c_2 e^{\lambda x} = 0. \tag{4}$$

If we cancel the (nonzero) exponentials we obtain a quadratic equation for λ, solution of which gives λ as a function of x. However, λ was supposed to be a constant, so we have a contradiction, and the method fails. (Specifically, if λ turns out to be a function of x, then $y' = \lambda e^{\lambda x}$ and $y'' = \lambda^2 e^{\lambda x}$, above, were incorrect.) Said differently, the $x^2 e^{\lambda x}, x e^{\lambda x}, e^{\lambda x}$ terms in (4) are LI and cannot be made to cancel identically to zero on any given x interval.

The reason we have discussed this fruitless approach is to emphasize that it is incorrect, and to caution against using it. By contrast, if the equation were of constant-coefficient type, say $y'' + c_1 y' + c_2 y = 0$, then $y = e^{\lambda x}$ would work because $y = e^{\lambda x}$, $y' = \lambda e^{\lambda x}$, $y'' = \lambda^2 e^{\lambda x}$ are LD, so the combination $y'' + c_1 y' + c_2 y$ could be made to cancel to zero by suitable choice of λ.

Although the form $e^{\lambda x}$ will not work for Cauchy–Euler equations, the form

$$\boxed{y = x^\lambda} \tag{5}$$

will, because $y = x^\lambda, xy' = \lambda x^\lambda, x^2 y'' = \lambda(\lambda - 1)x^\lambda, \ldots$ are LD since each is a constant times x^λ. Putting (5) into the second-order equation (3) gives

$$[\lambda(\lambda - 1) + c_1 \lambda + c_2]\, x^\lambda = 0.$$

Since $x^\lambda \neq 0$, we require of λ that

$$\lambda^2 - (1 - c_1)\lambda + c_2 = 0,$$

so

$$\lambda = \frac{1 - c_1 \pm \sqrt{(1 - c_1)^2 - 4c_2}}{2}. \tag{6}$$

We distinguish three cases, depending upon whether the discriminant $\Delta \equiv (1 - c_1)^2 - 4c_2$ is positive, zero, or negative:

$\Delta > 0$: **Distinct real roots.** If $\Delta > 0$, then (6) gives two distinct real roots, say λ_1 and λ_2, so we have the general solution to (3) as

$$\boxed{y(x) = Ax^{\lambda_1} + Bx^{\lambda_2}.} \tag{7}$$

EXAMPLE 1. To solve

$$x^2y'' - 2xy' - 10y = 0, \tag{8}$$

seek $y = x^\lambda$. That form gives $\lambda^2 - 3\lambda - 10 = 0$, with roots $\lambda = -2$ and 5, so the general solution of (8) is

$$y(x) = \frac{A}{x^2} + Bx^5. \quad ■$$

$\Delta = 0$: **Repeated real roots.** In this case (6) gives the repeated root $\lambda = \dfrac{1 - c_1}{2} \equiv \lambda_1$. Thus we have the solution Ax^{λ_1}, but are missing a second linearly independent solution, which is needed if we are to obtain a general solution of (3). Evidently, the missing solution is not of the form x^λ, or we would have found it when we sought $y(x) = x^\lambda$.

To find the missing solution we use Lagrange's method of **reduction of order**, as we did in Section 3.3.3 for constant-coefficient differential equations with repeated roots of the characteristic equation. That is, we let A vary, and seek

$$y(x) = A(x)x^{\lambda_1}. \tag{9}$$

Putting (9) into (3) gives (we leave the details to the reader, as Exercise 3)

$$xA'' + A' = 0.$$

Next, set $A' \equiv p$, say, to reduce the order:

$$x\frac{dp}{dx} + p = 0, \tag{10}$$

so $p = D/x$ and $A(x) = D \ln x + C$, where C, D are arbitrary constants (Exercise 4). Finally, putting the latter back into (9) gives the general solution of (3) as

$$\boxed{y(x) = (C + D \ln x)\, x^{\lambda_1}.} \tag{11}$$

EXAMPLE 2. To solve

$$x^2y'' + 7xy' + 9y = 0, \tag{12}$$

seek $y = x^\lambda$. That form gives $\lambda^2 + 6\lambda + 9 = 0$, with the repeated root $\lambda = -3$, so the general solution of (12) is

$$y(x) = (A + B \ln x)\, x^{-3}. \quad ∎$$

$\Delta < 0$: **Complex roots.** In this case (6) gives the distinct complex-conjugate roots

$$\lambda = \frac{1 - c_1}{2} \pm i \frac{\sqrt{4c_2 - (1 - c_1)^2}}{2} \equiv \alpha \pm i\beta, \tag{13}$$

so we have the general solution of (3) as

$$\begin{aligned} y(x) &= A x^{\alpha + i\beta} + B x^{\alpha - i\beta} \\ &= x^\alpha \left(A x^{i\beta} + B x^{-i\beta} \right). \end{aligned} \tag{14}$$

However, since we normally prefer real solution forms, let us use the identity $u = e^{\ln u}$ to re-express (14) as*

$$\begin{aligned} y(x) &= x^\alpha \left(A e^{\ln x^{i\beta}} + B e^{\ln x^{-i\beta}} \right) = x^\alpha \left(A e^{i\beta \ln x} + B e^{-i\beta \ln x} \right) \\ &= x^\alpha \left\{ A \left[\cos(\beta \ln x) + i \sin(\beta \ln x) \right] + B \left[\cos(\beta \ln x) - i \sin(\beta \ln x) \right] \right\} \\ &= x^\alpha \left[(A + B) \cos(\beta \ln x) + i(A - B) \sin(\beta \ln x) \right]. \end{aligned} \tag{15}$$

Or, letting $A + B \equiv C$ and $i(A - B) \equiv D$,

$$\boxed{y(x) = x^\alpha \left[C \cos(\beta \ln x) + D \sin(\beta \ln x) \right].} \tag{16}$$

EXAMPLE 3. To solve

$$x^2 y'' - 2xy' + 4y = 0, \tag{17}$$

*It is important to appreciate that the $x^{\alpha \pm i\beta}$ quantities, in (14), are "new objects" for us, for we have not yet (in this book) defined a real number x raised to a complex power (unless x happens to be e, in which case the already-discussed Euler's formulas apply). Staying as close as possible to familiar *real* variable results, let us write

$$x^{\alpha + i\beta} = x^\alpha x^{i\beta} = x^\alpha \left(e^{\ln x^{i\beta}} \right) = x^\alpha e^{i\beta \ln x},$$

and similarly for $x^{\alpha - i\beta}$. None of these three equalities are justifiable, since they rely on the formulas $x^{a+b} = x^a x^b, u = e^{\ln u}$, and $\ln x^c = c \ln x$, which assume x, a, b, c to be real, but we hereby understand them to hold by definition. Observe that complex quantities and complex functions keep forcing themselves upon us. Therefore, it behooves us to establish a general theory of complex functions, rather than deal with these issues one by one. We will do exactly that, but not until much later in the text.

seek $y = x^\lambda$. That form gives $\lambda^2 - 3\lambda + 4 = 0$, so $\lambda = \frac{3}{2} \pm i\frac{\sqrt{7}}{2}$. Hence

$$\begin{aligned} y(x) &= Ax^{3/2+i\sqrt{7}/2} + Bx^{3/2-i\sqrt{7}/2} = x^{3/2}\left(Ax^{i\sqrt{7}/2} + Bx^{-i\sqrt{7}/2}\right) \\ &= x^{3/2}\left(Ae^{i\frac{\sqrt{7}}{2}\ln x} + Be^{-i\frac{\sqrt{7}}{2}\ln x}\right) \\ &= x^{3/2}\left[C\cos\left(\frac{\sqrt{7}}{2}\ln x\right) + D\sin\left(\frac{\sqrt{7}}{2}\ln x\right)\right]. \quad \blacksquare \end{aligned}$$

Recall that we have limited our discussion of (3) to the case where $\mathbf{x} > \mathbf{0}$. The reason for that limitation is as follows. For a function $y(x)$ to be a solution of (3) on an x interval I, we first of all need each of y, y', and y'' to exist on I; that is, to be defined there. The function $\ln x$ and its derivatives are not defined at $x = 0$, nor is $\ln x$ defined (as a real-valued function) for $x < 0$. The functions $x^{\lambda_1}, x^{\lambda_2}$ in (7), x^{λ_1} in (11), and x^α in (16) cause similar problems. To deal with the case where $\mathbf{x} < \mathbf{0}$, it is more convenient to make the change of variable $x = -\xi$ in (3), so that ξ will be positive. Letting $y(x) = y(-\xi) \equiv Y(\xi)$,*

$$\frac{dy}{dx} = \frac{dY}{d\xi}\frac{d\xi}{dx} = -\frac{dY}{d\xi}, \qquad \frac{d^2y}{dx^2} = \frac{d}{d\xi}\left(-\frac{dY}{d\xi}\right)\frac{d\xi}{dx} = \frac{d^2Y}{d\xi^2}, \tag{18}$$

so (3) becomes

$$\xi^2\frac{d^2Y}{d\xi^2} + c_1\xi\frac{dY}{d\xi} + c_2Y = 0, \qquad (\xi > 0) \tag{19}$$

which is the same as (3)! Thus, its solutions are the same, but with x changed to ξ. For the case of distinct real roots, for instance,

$$y(x) = Ax^{\lambda_1} + Bx^{\lambda_2}$$

for $x > 0$, and

$$y(x) = Y(\xi) = Y(-x) = A(-x)^{\lambda_1} + B(-x)^{\lambda_2}$$

for $x < 0$. Observe that both of these forms are accounted for by the single expression

$$y(x) = A\,|x|^{\lambda_1} + B\,|x|^{\lambda_2}.$$

Similarly for the other cases (repeated real roots and complex roots). Let us state these results, for reference, as a theorem.

*Why do we change the name of the dependent variable from y to Y? Because they are different functions. To illustrate, suppose $y(x) = 5 + x^3$. Then $Y(\xi) = 5 + (-\xi)^3 = 5 - \xi^3$. For instance, if the argument of y is 2, then y is 13, but when the argument of Y is 2, then Y is -3.

THEOREM 3.6.1 *Second-Order Cauchy–Euler Equation*
The general solution of the second-order Cauchy–Euler equation

$$x^2y'' + c_1xy' + c_2y = 0, \tag{20}$$

on any x interval not containing the origin, is

$$y(x) = \begin{cases} A\,|x|^{\lambda_1} + B\,|x|^{\lambda_2} \\ (A + B\ln|x|)\,|x|^{\lambda_1} \\ |x|^{\alpha}\,[A\cos(\beta\ln|x|) + B\sin(\beta\ln|x|)] \end{cases} \tag{21}$$

if the roots λ_1, λ_2 of $\lambda^2 + (c_1 - 1)\lambda + c_2 = 0$ are real and distinct, real and repeated, or complex ($\lambda = \alpha \pm i\beta$), respectively.

Of course, if the x interval is to the right of the origin, then the absolute value signs in (21) can be dropped.

To close our discussion of the Cauchy–Euler equation, consider the higher-order case, $n > 2$. For simplicity, we consider $x > 0$; as for the second-order case treated above, $x < 0$ can be handled simply by changing all x's in the solution to $|x|$.

EXAMPLE 4. Consider the third-order Cauchy–Euler equation

$$x^3y''' - 3x^2y'' + 7xy' - 8y = 0. \tag{22}$$

Seeking $y(x) = x^{\lambda}$ gives

$$\lambda^3 - 6\lambda^2 + 12\lambda - 8 = 0,$$

with the roots $\lambda = 2, 2, 2$. Thus we have the solution $y(x) = Ax^2$, but we need two more linearly independent solutions. To find them, we use reduction of order and seek $y(x) = A(x)x^2$. Putting that form into (22) gives the equation

$$x^2A''' + 3xA'' + A' = 0$$

on $A(x)$, which can be reduced to the second-order equation

$$x^2p'' + 3xp' + p = 0 \tag{23}$$

by letting $A' = p$. The latter is again of Cauchy–Euler type, and letting $p(x) = x^{\lambda}$ gives $\lambda = -1, -1$, so that

$$p(x) = (B + C\ln x)\frac{1}{x}.$$

Since $A' = p$,

$$A(x) = \int p\,dx = B\ln x + C\frac{(\ln x)^2}{2} + D,$$

and

$$y(x) = \left[C_1 + C_2 \ln x + C_3(\ln x)^2\right] x^2 \tag{24}$$

is the desired general solution of (22). ▮

Comparing the latter result with the solution (11) for the second-order Cauchy–Euler equation with repeated root λ_1, we might well suspect that if any Cauchy–Euler equation has a repeated root λ_1 of order k, then that root contributes the form

$$\left[C_1 + C_2 \ln x + C_3(\ln x)^2 + \cdots + C_k(\ln x)^{k-1}\right] x^{\lambda_1} \tag{25}$$

to the general solution. We state, without proof, that that is indeed the case.

EXAMPLE 5. As a summary example, suppose that upon seeking solutions of a given eighth-order Cauchy–Euler equation in the form $y(x) = x^\lambda$ we obtain the roots

$$\lambda = -2.4,\ 1.7,\ 1.7,\ 1.7,\ -3+4i,\ -3+4i,\ -3-4i,\ -3-4i.$$

Then the general solution is

$$\begin{aligned} y(x) = {} & C_1 x^{-2.4} + \left[C_2 + C_3(\ln x) + C_4(\ln x)^2\right] x^{1.7} \\ & + (C_5 + C_6 \ln x)\, x^{-3+4i} + (C_7 + C_8 \ln x)\, x^{-3-4i}, \end{aligned} \tag{26}$$

or (Exercise 5),

$$\begin{aligned} y(x) = {} & C_1 x^{-2.4} + \left[C_2 + C_3(\ln x) + C_4(\ln x)^2\right] x^{1.7} \\ & + \{[C_9 \cos(4\ln x) + C_{10} \sin(4\ln x)] \\ & \quad + \ln x\, [C_{11} \cos(4\ln x) + C_{12} \sin(4\ln x)]\}\, x^{-3}. \end{aligned} \tag{27}$$

Although such high-order Cauchy–Euler equations are uncommon, we include this example to illustrate the general solution pattern for any order $n \geq 2$. ▮

This concludes our discussion of the Cauchy–Euler equation. We will meet Cauchy–Euler equations again in the next chapter, in connection with power series solutions of differential equations, and again when we study the partial differential equations governing such phenomena as heat conduction, electric potential, and certain types of fluid flow, in later chapters.

3.6.2. Reduction of order. (Optional) We have already used Lagrange's method of reduction of order to find "missing solutions," for constant-coefficient equations and for Cauchy–Euler equations as well. Here, we focus not on constant-coefficient or Cauchy–Euler equations, but on the method itself and indicate its more general application to any linear homogeneous differential equation.

For definiteness, consider the second-order case,

$$y'' + a_1(x)y' + a_2(x)y = 0. \tag{28}$$

Suppose that one solution is known to us, say $Y(x)$, and that a second linearly independent solution is sought. If $Y(x)$ is a solution, then so is $AY(x)$ for any constant A. The idea behind Lagrange's method of **reduction of order** is to seek the missing solution in the form

$$y(x) = A(x)Y(x), \tag{29}$$

where $A(x)$ is to be determined.

The method is similar to Lagrange's method of *variation of parameters*, introduced in Section 2.2, but its purpose is different. The latter was used to find the general solution of the *non*homogeneous equation $y' + p(x)y = q(x)$ from the solution $y_h(x) = Ae^{-\int p(x)\,dx}$ of the homogeneous equation $y' + p(x)y = 0$, by varying the parameter A and seeking $y(x) = A(x)e^{-\int p(x)\,dx}$. Reduction of order is similar in that we vary the parameter A in $y = AY(x)$, but different in that it is used to find a missing solution of a homogeneous equation from a known solution $Y(x)$ of that homogeneous equation.

We begin by emphasizing that at first glance the form (29) seems to be without promise. To explain that statement, observe that the search for a pair of lost glasses can be expected to be long and arduous if we merely know that they are somewhere in North America, but shorter and easier to whatever extent we are able to narrow the domain of the search. If, for instance, we know that they are somewhere on our desk, then the search is short and simple. Likewise, when we solve a constant-coefficient equation by seeking y in the form $e^{\lambda x}$ then the search is short and simple since, first, solutions will indeed be found within that subset and, second, because that subset is tiny compared to the set of all possible functions, just as one's desk is tiny compared to North America. Similarly, when we solve a Cauchy–Euler equation by seeking y in the form x^{λ}.

With this idea in mind, observe that the form (29) does not narrow our search in the slightest, since it includes *all* functions! That is, any given function $f(x)$ can be expressed as $A(x)Y(x)$ simply by choosing $A(x)$ to be $f(x)/Y(x)$.

Proceeding nonetheless, we put (29) into (28) and obtain the differential equation

$$A''Y + (2Y' + a_1Y)A' + (Y'' + a_1Y' + a_2Y)A = 0 \tag{30}$$

on $A(x)$. At first glance it appears that this differential equation on $A(x)$ is probably even harder than the original equation, (28), on $y(x)$. However, and this is the heart of Lagrange's idea, all of the undifferentiated A terms must cancel, because if A were a constant (in which case the A' and A'' terms would all drop out), then the remaining terms would have to cancel to zero because $AY(x)$ is a solution of (28). Thus, the coefficient of A in (30) is zero, so (30) becomes

$$A''Y + (2Y' + a_1Y)A' = 0, \tag{31}$$

the order of which can now be reduced from two to one by letting $A' = p$:

$$\frac{dp}{dx} + \left(\frac{2Y' + a_1Y}{Y}\right)p = 0. \tag{32}$$

Integrating the latter gives

$$\begin{aligned} p(x) &= Be^{-\int \frac{2Y'+a_1Y}{Y}\,dx} = Be^{-2\int \frac{dY}{Y} - \int a_1\,dx} \\ &= BY(x)^{-2}e^{-\int a_1(x)\,dx}. \end{aligned}$$

Finally, integration of $A' = p$ gives

$$A(x) = \int p(x)\,dx = B\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx + C,$$

so (29) becomes

$$y(x) = \left[B\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx + C\right]Y(x). \tag{33}$$

The $CY(x)$ term merely reproduces the solution that was already known; the missing solution is provided by the other term, $BY(x)\int Y(x)^{-2}e^{-\int a_1(x)\,dx}dx$. That this solution and the original solution $Y(x)$ are necessarily LI is left for Exercise 6.

Incidentally, the result (33) could also be written using definite integrals if one prefers, as

$$y(x) = \left[B\int_\alpha^x Y(\xi)^{-2}e^{-\int_\beta^\xi a_1(\eta)\,d\eta}d\xi + C\right]Y(x), \tag{34}$$

where the lower limits α and β are arbitrary numbers, for the effect of changing α is simply to add some constant to the ξ integral, and that constant times B can be absorbed by the arbitrary constant C. Likewise, changing β simply adds some constant, say P, to the η integral, and the resulting e^{-P} factor can be absorbed by the arbitrary constant B.

EXAMPLE 6. *Legendre's equation.* The equation

$$(1-x^2)y'' - 2xy' + 2y = 0, \qquad (-1 < x < 1) \tag{35}$$

is known as **Legendre's equation**, after the French mathematician *Adrien Marie Legendre* (1752–1833). It is studied in Chapter 4, and used in later chapters when it arises in physical applications.

Observe that (35) admits the simple solution $y(x) = x$. To find a second solution, and hence a general solution, we can seek $y(x) = A(x)x$ and follow the steps outlined above. Rather, let us simply use the derived result (33). First, we divide (35) by $1 - x^2$ to reduce it to the form (28), so that we can identify $a_1(x)$. Thus, with $a_1(x) = -\dfrac{2x}{1-x^2}$ and $Y(x) = x$, (33) gives

$$\begin{aligned} y(x) &= \left[B\int \frac{e^{\int 2x\,dx/(1-x^2)}}{x^2}dx + C\right]x = \left[B\int \frac{dx}{x^2(1-x^2)} + C\right]x \\ &= \left[B\int\left(\frac{1}{x^2} + \frac{1}{2}\frac{1}{1-x} + \frac{1}{2}\frac{1}{1+x}\right)dx + C\right]x \\ &= \left[B\left(-\frac{1}{x} + \frac{1}{2}\ln\frac{1+x}{1-x}\right) + C\right]x, \end{aligned}$$

or, equivalently,

$$y(x) = C_1 x + C_2 \left(1 - \frac{x}{2} \ln \frac{1+x}{1-x}\right). \quad ▮$$

In this example we were able to evaluate the integrals that occurred. In other cases we may not be able to, even with the help of computer software, and may therefore need to leave the answer in integral form.

3.6.3. Factoring the operator. (Optional) We have been considering the nth-order linear homogeneous equation

$$L[y] = \left[\frac{d^n}{dx^n} + a_1(x)\frac{d^{n-1}}{dx^{n-1}} + \cdots + a_n(x)\right] y = 0,$$

or

$$\left(D^n + a_1 D^{n-1} + \cdots + a_n\right) y = 0, \tag{36}$$

where $D = \dfrac{d}{dx}$, $D^2 = DD = \dfrac{d}{dx}\dfrac{d}{dx} = \dfrac{d^2}{dx^2}$, and so on.

Suppose, first, that (36) is of constant-coefficient type (i.e., each a_j is a constant), and that the characteristic polynomial $\lambda^n + a_1\lambda^{n-1} + \cdots + a_n$ can be factored as $(\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n)$, where one or more of the roots λ_j may be repeated. Then the differential operator $L = D^n + a_1 D^{n-1} + \cdots + a_n$ can be factored in precisely the same way, as $(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)$, where we understand $(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)y$ to mean that first y is operated on by $D - \lambda_n$, then the result of that step is operated on by $D - \lambda_{n-1}$, and so on. That is, we begin at the right and proceed to the left. Further, it is readily verified that the sequential order of the $D - \lambda_j$ factors is immaterial, that is, they commute. If $n = 2$, for instance,

$$\begin{aligned}(D - \lambda_1)(D - \lambda_2)y &= (D - \lambda_1)(y' - \lambda_2 y)\\ &= D(y' - \lambda_2 y) - \lambda_1(y' - \lambda_2 y)\\ &= y'' - (\lambda_2 + \lambda_1)y' + \lambda_1\lambda_2 y\end{aligned}$$

and

$$\begin{aligned}(D - \lambda_2)(D - \lambda_1)y &= (D - \lambda_2)(y' - \lambda_2 y)\\ &= D(y' - \lambda_1 y) - \lambda_2(y' - \lambda_1 y)\\ &= y'' - (\lambda_2 + \lambda_1)y' + \lambda_2\lambda_1 y\end{aligned}$$

are the same.

By factoring L, we are able to reduce the solution of

$$(D - \lambda_1)(D - \lambda_2)\cdots(D - \lambda_n)y = 0 \tag{37}$$

to the solution of a sequence of n first-order equations, each of which is of the form $y' - py = q$ or

$$(D - p)y = q, \tag{38}$$

where p is a constant and $q(x)$ is known. From Section 2.2, we know that the solution of (38) is

$$y(x) = e^{px}\left(\int e^{-px}q(x)\,dx + A\right), \tag{39}$$

where A is an arbitrary constant.

Let us illustrate with an example.

EXAMPLE 7. The equation

$$y''' - 3y'' + 4y = 0 \tag{40}$$

admits the characteristic roots $\lambda = -1, 2, 2$, so we can factor (40) as

$$(D + 1)(D - 2)(D - 2)y = 0. \tag{41}$$

We begin the solution procedure by setting

$$(D - 2)(D - 2)y = u, \tag{42}$$

so that (41) becomes

$$(D + 1)u = 0,$$

with the solution

$$u(x) = Ae^{-x}.$$

Putting the latter into (42) gives

$$(D - 2)(D - 2)y = Ae^{-x}, \tag{43}$$

in which we set

$$(D - 2)y = v. \tag{44}$$

Then (43) becomes

$$(D - 2)v = Ae^{-x},$$

with the solution

$$v(x) = e^{2x}\left(\int e^{-2x}Ae^{-x}dx + B\right) = -\frac{A}{3}e^{-x} + Be^{2x}.$$

Finally, putting the latter into (44) gives

$$(D - 2)y = -\frac{A}{3}e^{-x} + Be^{2x},$$

with the solution

$$\begin{aligned} y(x) &= e^{2x}\left[\int e^{-2x}\left(-\frac{A}{3}e^{-x} + Be^{2x}\right)dx + C\right] \\ &= \frac{A}{9}e^{-x} + Bxe^{2x} + Ce^{2x}, \end{aligned}$$

or, equivalently,

$$y(x) = C_1e^{-x} + (C_2 + C_3x)\,e^{2x},$$

which is the same solution as obtained by methods discussed in earlier sections. Notice, in particular, that the presence of the repeated root $\lambda = 2, 2$ presented no additional difficulty. ■

Although the factorization method reduces an nth-order equation to a sequence of n first-order equations, it is quite different from the method of reduction of order described above in Section 3.6.2.

Thus far we have limited our discussion of factorization to the constant-coefficient case. The nonconstant-coefficient case is more difficult. To appreciate the difficulty, consider the equation

$$y'' - x^2y = (D^2 - x^2)y = 0. \tag{45}$$

If we can factor $D^2 - x^2$ as $(D+x)(D-x)$, then we can solve (45) by the method outlined above. However,

$$\begin{aligned} (\underline{D} + x)(D - \underline{x})y &= (D + x)(y' - xy) = D(y' - xy) + x(y' - xy) \\ &= y'' - xy' - y + xy' - x^2y = y'' - (x^2 + 1)y, \end{aligned} \tag{46}$$

so $(D+x)(D-x) = D^2 - (x^2+1)$ is not the same as $D^2 - x^2$. The problem is that the differential operator on the left-hand side of (46) acts not only on y but also on itself, in the sense that an additional term is contributed to the final result, namely $-y$, through the action of the underlined D on the underlined x. Observe, further, that $D + x$ and $D - x$ do not commute since $(D + x)(D - x) = D^2 - (x^2 + 1)$, whereas $(D - x)(D + x) = D^2 - (x^2 - 1)$.

Thus, the following practical question arises: given a nonconstant coefficient operator, can it be factored and, if so, how?

Limiting our attention to equations of second order (which, arguably, is the most important case in applications), suppose that $a_1(x)$ and $a_2(x)$ are given, and that we seek $a(x)$ and $b(x)$ so that

$$y'' + a_1(x)y' + a_2(x)y = [D - a(x)][D - b(x)]y. \tag{47}$$

Writing out the right-hand side,

$$\begin{aligned} y'' + a_1y' + a_2y &= (D - a)(y' - by) \\ &= y'' - (a + b)y' + (ab - b')y. \end{aligned} \tag{48}$$

Since this equation needs to hold for all (twice-differentiable) functions y, a and b must satisfy the conditions (Exercise 13)

$$a + b = -a_1, \tag{49a}$$
$$ab - b' = a_2, \tag{49b}$$

or, isolating a and b (Exercise 14),

$$a' = a^2 + (a_1)a + (a_2 - a_1'), \tag{50a}$$
$$b' = -b^2 - (a_1)b - (a_2). \tag{50b}$$

Each of these equations is a special case of the nonlinear **Riccati equation**

$$\boxed{y' = p(x)y^2 + q(x)y + r(x),} \tag{51}$$

which was discussed in Exercise 11 of Section 2.2.

Thus, from a global point of view, it is interesting to observe that *the class of second-order equations with nonconstant coefficients is, in a sense, equivalent in difficulty to the class of nonlinear first-order Riccati equations.* We saw, in Exercise 11 of Section 2.2 that in the exceptional case where a particular solution $Y(x)$ of (51) can be found, perhaps by inspection, the nonlinear equation (51) can be converted to the linear equation

$$v' + [2p(x)Y(x) + q(x)]\, v = -p(x) \tag{52}$$

by the change of variables

$$y = Y(x) + \frac{1}{v}. \tag{53}$$

Thus, just as we are able to solve the Riccati equation only in exceptional cases, we are able to factor second-order nonconstant coefficient equations (and solve them readily) only in exceptional cases. In general, then, nonconstant-coefficient differential equations are hard in that we are unable to find closed form solutions.

EXAMPLE 8. Consider the equation

$$y'' - (x^2 + 1)y = 0. \tag{54}$$

Here $a_1(x) = 0$ and $a_2(x) = -(x^2 + 1)$, so (50a,b) are

$$a' = a^2 - x^2 - 1, \tag{55a}$$
$$b' = -b^2 + x^2 + 1. \tag{55b}$$

In this case we are lucky enough to notice the particular solution $a(x) = -x$ of (55a). Putting this result into (49a) then gives $b(x) = x$. [Equivalently, we could have noticed the particular solution $b(x) = x$ of (55b) and then obtained $a(x) = -x$ from (49a).] Thus, we have the factorization

$$y'' - (x^2 + 1)y = (D + x)(D - x)y = 0. \tag{56}$$

Proceeding as outlined above, we are able (Exercise 15) to derive the general solution

$$y(x) = Ae^{x^2/2} + Be^{x^2/2} \int e^{-x^2}\,dx. \tag{57}$$

Going one step further, suppose that initial conditions $y(0) = 0$ and $y'(0) = 1$ are prescribed and that we wish to evaluate A and B. First, we re-express (57) in the equivalent and more convenient form

$$y(x) = Ae^{x^2/2} + Be^{x^2/2} \int_0^x e^{-\xi^2}\,d\xi. \tag{58}$$

We could have used any lower integration limit, but 0 will be most convenient because the initial conditions are at $x = 0$. Then

$$\begin{aligned} y(0) &= 0 = (A)(1) + (B)(0), \\ y'(0) &= 1 = (A)(0) + (B)(1), \end{aligned}$$

where we have used the fundamental theorem of the calculus (Section 2.2) to differentiate the integral term. Thus, $A = 0$ and $B = 1$, so

$$y(x) = e^{x^2/2} \int_0^x e^{-\xi^2}\,d\xi \tag{59}$$

is the desired particular solution. ■

The integral in (59) is nonelementary in that it cannot be evaluated in closed form in terms of the elementary functions. But it arises often enough so that it has been used to define a new function, the so-called **error function**

$$\boxed{\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\xi^2}\,d\xi,} \tag{60}$$

where the $2/\sqrt{\pi}$ is included to normalize $\text{erf}(x)$ to unity as $x \to \infty$ since (as will be shown in a later chapter)

$$\int_0^\infty e^{-\xi^2}\,d\xi = \frac{\sqrt{\pi}}{2}. \tag{61}$$

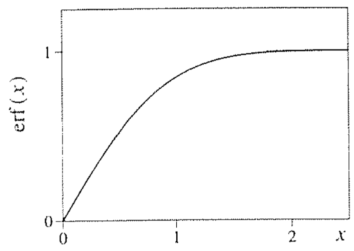


Figure 1. The error function $\text{erf}(x)$.

The graph of $\text{erf}(x)$ is shown in Fig. 1 for $x > 0$. For $x < 0$ we rely on the fact that $\text{erf}(-x) = -\text{erf}(x)$ (Exercise 18); for instance, $\text{erf}(-\infty) = -\text{erf}(\infty) = -1$. Since $e^{-\xi^2}$ is (to within a scale factor) a *normal probability distribution*, one way in which the error function arises is in the study of phenomena that are governed by normal distributions. For instance, we will encounter the error function when we study the movement of heat by the physical process of conduction.

Thus, our solution (59) can be re-expressed as $y(x) = \sqrt{\pi/2}\,e^{x^2/2}\,\text{erf}(x)$. Just as we know the values of $\sin x$, its Taylor series, and its various properties, likewise we know the values of $\text{erf}(x)$, its Taylor series, and its various properties, so we should feel comfortable with $\text{erf}(x)$ and regard it henceforth as a known function. Though not included among the so-called "elementary functions," it is one of many "special functions" that are now available in the engineering science and mathematics literature.

Closure. We have seen, in this section, that nonconstant-coefficient equations can be solved in closed form only in exceptional cases. The most important of these is the case of the Cauchy–Euler equation

$$x^n \frac{d^n y}{dx^n} + c_1 x^{n-1} \frac{d^{n-1} y}{dx^{n-1}} + \cdots + c_{n-1} x \frac{dy}{dx} + c_n y = 0.$$

Recall that a constant-coefficient equation necessarily admits at least one solution in the form $e^{\lambda x}$, and that in the case of a repeated root of order k the solutions corresponding to that root can be found by reduction of order to be $\left(C_1 + C_2 x + \cdots + C_k x^{k-1}\right) e^{\lambda x}$. Analogously, a Cauchy–Euler equation necessarily admits at least one solution in the form x^λ, and in the case of a repeated root of order k the solutions corresponding to that root can be found by reduction of order to be $\left[C_1 + C_2 \ln x + \cdots + C_k (\ln x)^{k-1}\right] x^\lambda$.

In fact, it turns out that the connection between constant-coefficient equations and Cauchy – Euler equations is even closer than that in as much as *any given Cauchy – Euler equation can be reduced to a constant-coefficient equation* by a change of independent variable according to $x = e^t$. Discussion of that point is reserved for the exercises.

Beyond our striking success with the Cauchy–Euler equation, other successes for nonconstant-coefficient equations are few and far between. For instance, we might be able to obtain one solution by inspection and others, from it, by reduction of order. Or, we might, in exceptional cases, be successful in factoring the differential operator but, again, such successes are exceptional. Thus, other lines of approach will be needed for nonconstant-coefficient equations, and they are developed in Chapters 4 and 6.

EXERCISES 3.6

1. Derive a general solution for the given Cauchy–Euler equation by seeking $y(x) = x^\lambda$. That is, derive the solution, rather than merely use any stated result such as Theorem 3.6.1. In addition, find the particular solution corresponding to the initial conditions, if such conditions are given, and state the interval of validity of that solution.

(a) $xy' + y = 0$
(b) $xy' - y = 0; \quad y(2) = 5$
(c) $xy'' + y' = 0$
(d) $xy'' - 4y' = 0; \quad y(1) = 0, \quad y'(1) = 3$
(e) $x^2y'' + xy' - 9y = 0; \quad y(2) = 1, \quad y'(2) = 2$
(f) $x^2y'' + xy' + y = 0; \quad y(1) = 1, \quad y'(1) = 0$
(g) $x^2y'' + 3xy' + 2y = 0; \quad y(1) = 0, \quad y'(1) = 2$
(h) $x^2y'' - 2y = 0; \quad y(-5) = 3, \quad y'(-5) = 0$
(i) $4x^2y'' + 5y = 0; \quad y(1) = 0, \quad y'(1) = 1$
(j) $x^2y'' + xy' + 4y = 0$
(k) $x^2y'' + 2xy' - 2y = 0; \quad y(3) = 2, \quad y'(3) = 2$
(l) $(x+2)^2y'' - y = 0$ HINT: Let $x + 2 = t$.
(m) $x^2y''' - 2y' = 0; \quad y(1) = 2, \quad y'(1) = y''(1) = 0$
(n) $xy''' - y'' = 0; \quad y(0) = 1, \quad y'(0) = y''(0) = 0$
(o) $x^2y'' + xy' - \kappa^2 y = 0$ (κ a constant)
(p) $x^3y''' + xy' - y = 0$

(q) $x^3y''' + 2xy' - 2y = 0$
(r) $x^2y''' + xy'' - y' = 0$
(s) $x^3y''' + 6x^2y'' + 7xy' + y = 0$
(t) $x^4y'''' + 6x^3y''' + 3x^2y'' - 3xy' + 4y = 0$
(u) $x^4y'''' + 6x^3y''' + 3x^2y'' - 3xy' + y = 0$
(v) $xy'''' - 2y''' = 0; \quad y(1) = 5,$
$y'(1) = y''(1) = y'''(1) = 0$

2. (a)–(v) For the corresponding problem in Exercise 1, use computer software to obtain a general solution, as well as a particular solution if initial conditions are given.

3. Putting (9) into (3), show that the equation $xA'' + A' = 0$ results, as claimed below equation (9).

4. Solve (10), and derive the general solution $A(x) = D \ln x + C$ stated below equation (10).

5. Fill in the steps between (26) and (27).

6. Prove that the two solutions within (33) [or, equivalently, (34)] are necessarily LI. You may assume that $a_1(x)$ is continuous on the x interval of interest. HINT: Recall the fundamental theorem of the calculus (given in Section 2.2).

7. It was stated in the Closure that *any given Cauchy–Euler equation can be reduced to a constant-coefficient equation by the change of variables* $x = e^t$. In this exercise we ask you to try that idea for some specific cases; in the next exercise we ask for a general proof of the italicized claim. Let $y(x(t)) \equiv Y(t)$, and let y' and Y' denote dy/dx and dY/dt, respectively.

(a) Show that the change of variables $x = e^t$ reduces the Cauchy–Euler equation $x^2y'' - xy' - 3y = 0$ to the constant-coefficient equation $Y'' - 2Y' - 3Y = 0$. Thus, show that $Y(t) = Ae^{-t} + Be^{3t}$. Since $t = \ln x$, show that $y(x) = Ax^{-1} + Bx^3$.
(b) Same as (a), for $x^2y'' + xy' - 4y = 0$.
(c) Same as (a), for $x^2y'' + xy' + 4y = 0$.
(d) Same as (a), for $x^2y'' + 3xy' + y = 0$.
(e) Same as (a), for $x^2y'' + xy' - 9y = 0$.
(f) Same as (a), for $x^2y'' + y = 0$.
(g) Same as (a), for $x^2y'' + 2xy' - 2y = 0$.
(h) Same as (a), for $4x^2y'' - y = 0$.

8. First, read the introduction to Exercise 7. Consider the general Cauchy–Euler equation

$$\left(x^nD^n + c_1x^{n-1}D^{n-1} + \cdots + c_{n-1}xD + c_n\right)y = 0, \tag{8.1}$$

where $D = d/dx$. Let $x = e^t$, and define $y(x(t)) \equiv Y(t)$.

(a) Using chain differentiation, show that

$$\begin{aligned} xDy &= DY, \\ x^2D^2y &= D(D-1)Y, \\ x^3D^3y &= D(D-1)(D-2)Y, \end{aligned} \tag{8.2}$$

where D acting on $y(x)$ means d/dx and D acting on $Y(t)$ means d/dt.
(b) The results (8.2) suggest that the formula

$$x^kD^ky = D(D-1)\cdots(D-k+1)Y, \tag{8.3}$$

holds for all positive integers k. Prove that (8.3) is indeed correct. HINT: Use mathematical induction. That is, assume that (8.3) holds for any given positive k and, by differentiating both sides with respect to x, show that

$$x^{k+1}D^{k+1}y = D(D-1)\cdots(D-k)Y, \tag{8.4}$$

which is the same as (8.3) but with k changed to $k+1$. Thus, it must be true that (8.3) holds for all positive integers k.
(c) Finally, replacing each x^kD^ky in (8.1) by the corresponding right-hand side of (8.3), state why the resulting differential equation on $Y(t)$ will be of constant-coefficient type.

9. (*Electric potential*) The electric potential Φ within an annular region such as a long metal pipe of inner radius r_1 and outer radius r_2, satisfies the differential equation

$$r\frac{d^2\Phi}{dr^2} + \frac{d\Phi}{dr} = 0. \qquad (r_1 < r < r_2)$$

Solve for the potential distribution $\Phi(r)$ subject to these boundary conditions:

(a) $\Phi(r_1) = \Phi_1, \quad \Phi(r_2) = \Phi_2$
(b) $\dfrac{d\Phi}{dr}(r_1) = 0, \quad \Phi(r_2) = \Phi_2$

10. (*Steady-state temperature distribution*) The steady-state temperature distribution u within a hollow sphere, of inner radius r_1 and outer radius r_2, is governed by the differential equation

$$r\frac{d^2u}{dr^2} + 2\frac{du}{dr} = 0.$$

Solve for $u(r)$ subject to these boundary conditions:

(a) $u(r_1) = u_1, \quad u(r_2) = u_2$
(b) $\dfrac{du}{dr}(r_1) = 3, \quad u(r_2) = 0$
(c) $u(r_1) = u_1, \quad \dfrac{du}{dr}(r_2) = 0$

EXERCISES FOR OPTIONAL SECTION 3.6.2

11. Use the given solution $y_1(x)$ of the differential equation to find the general solution by the method of reduction of order (leaving the second solution in integral form if necessary).

(a) $y'' + xy' - y = 0; \quad y_1(x) = x$
(b) $xy'' + xy' - y = 0; \quad y_1(x) = x$
(c) $3xy'' - xy' + y = 0; \quad y_1(x) = x$
(d) $(x^2 - 1)y'' - 2y = 0; \quad y_1(x) = x^2 - 1$
(e) $2y'' + xy' - 2y = 0; \quad y_1(x) = x^2 + 2$

12. (a) – (e) Obtain a general solution of the corresponding differential equation in Exercise 11, using computer software.

EXERCISES FOR OPTIONAL SECTION 3.6.3

13. State, clearly and convincingly, logic by which (49a,b) follow from (48).

14. Fill in the steps between (49a,b) and (50a,b).

15. Provide the steps that are missing between the equation (56) and its solution (57).

16. If $a_1(x)$ and $a_2(x)$ are constants, then the factorization (47) should be simple. Show that the Riccati equations (50a,b) on a and b do indeed give the same results for a and b as can be obtained by more elementary means.

17. In general, the Riccati-type equations (50a,b) are hard. However, we should be able to solve them if the given nonconstant-coefficient equation $y'' + a_1(x)y' + a_2(x)y = 0$ is a Cauchy–Euler equation because that case is simple. Thus, use the method of factoring the operator for these equations:

(a) $x^2y'' - 2xy' + 2y = 0$
(b) $x^2y'' + xy' + 9y = 0$
(c) $x^2y'' + xy' - 9y = 0$
(d) $x^2y'' + 5xy' + 4y = 0$

18. From its integral definition, (60), show that $\text{erf}(-x) = -\text{erf}(x)$.

19. (*Integral representations*) The notion of an integral representation of a function, as used in (60) to define the error function $\text{erf}(x)$, might be unfamiliar to you. If so, it might help to point out that even the elementary functions can be introduced in that manner. For example, one can define the logarithm $\ln x$ as

$$\ln x = \int_1^x \frac{dt}{t}, \qquad (x > 0) \qquad (19.1)$$

from which formula the values of $\ln x$ can be derived (by numerical integration), and its various properties derived as well.

(a) To illustrate the latter claim, use (19.1) to derive the well known property $\ln x^a = a \ln x$ of the logarithm.
(b) Likewise, use (19.1) to derive the property $\ln(xy) = \ln x + \ln y$.

3.7 Solution of Nonhomogeneous Equation

Thus far, for differential equations of second-order and higher, we have studied only the homogeneous equation $L[y] = 0$. In this section we turn to the *non*homogeneous case

$$L[y] = f(x), \qquad (1)$$

where L is an nth-order linear differential operator. That is, this time we include a nonzero forcing function $f(x)$.

Before proceeding with solution techniques, let us reiterate that the function $f(x)$ (that is, what's left on the right-hand side when the terms involving y and its derivatives are put on the left-side side) is, essentially, a **forcing function**, and we will call it that, in this text, as a reminder of its physical significance.

For instance, we have already met (Section 1.3) the equation

$$m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx = F(t) \qquad (2)$$

governing the displacement $x(t)$ of a mechanical oscillator. Here, the forcing function is the applied force $F(t)$. For the analogous electrical oscillator (Section 2.3), governed by the equations

$$L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt} \tag{3}$$

on the current $i(t)$, and

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t) \tag{4}$$

on the charge $Q(t)$ on the capacitor, the forcing functions are the time derivative of the applied voltage $E(t)$, and the applied voltage $E(t)$, respectively.

As one more example, we give (without derivation) the differential equation

$$EI\frac{d^4y}{dx^4} + ky = w(x) \tag{5}$$

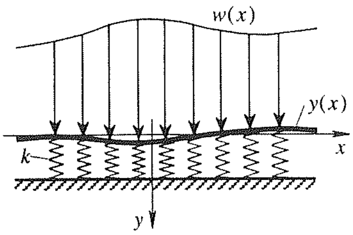


Figure 1. Beam on elastic foundation.

governing the deflection $y(x)$ of a beam that rests upon an elastic foundation, under a load distribution $w(x)$ (i.e., load per unit x length), as sketched in Fig. 1. E, I, and k are known physical constants: E is the Young's modulus of the beam material, I is the inertia of the beam's cross section, and k is the spring stiffness per unit length (force per unit length per unit length) of the foundation. Thus, in this case the forcing function is $w(x)$, the applied load distribution. [Derivation of (5) involves the so-called Euler beam theory and is part of a first course in solid mechanics.]

3.7.1. General solution. Remember that the general solution of the homogeneous equation $L[y] = 0$ is a family of solutions that contains every solution of that equation, over the interval of interest. Likewise, by the general solution of the nonhomogeneous equation $L[y] = f$, we mean a family of solutions that contains every solution of that equation, over the interval of interest.

Like virtually all of the concepts and methods developed in this chapter, the concepts that follow rest upon the assumed linearity of the differential equation (1), in particular, upon the fact that if L is linear then

$$L\left[\alpha u(x) + \beta v(x)\right] = \alpha L[u(x)] + \beta L[v(x)] \tag{6}$$

for any two functions u, v (n-times differentiable, of course, if L is an nth-order operator) and any constants α, β. Indeed, recall the analogous result for any number of functions:

$$L\left[\alpha_1 u_1(x) + \cdots + \alpha_k u_k(x)\right] = \alpha_1 L[u_1(x)] + \cdots + \alpha_k L[u_k(x)] \tag{7}$$

for any functions $u_1, \ldots, u_k$ and constants $\alpha_1, \ldots, \alpha_k$.

To begin, we suppose that $y_h(x)$ is a general solution of the homogeneous version of (1), $L[y] = 0$, and that $y_p(x)$ is any particular solution of (1): $L[y_p(x)] = f(x)$. That is, $y_p(x)$ is any function which, when put into the left-hand side of (1), gives $f(x)$. We will refer to $y_h(x)$ and $y_p(x)$ as **homogeneous** and **particular solutions** of (1), respectively. [Some authors write $y_c(x)$ in place of $y_h(x)$, and call it the **complementary solution**.]

THEOREM 3.7.1 *General Solution of* $L[y] = f$
If $y_h(x)$ and $y_p(x)$ are homogeneous and particular solutions of (1), respectively, on an interval I, then a general solution of (1), on I, is

$$y(x) = y_h(x) + y_p(x). \tag{8}$$

Proof: That (8) satisfies (1) follows from the linearity of (1):

$$\begin{aligned} L\left[y_h(x) + y_p(x)\right] &= L\left[y_h(x)\right] + L\left[y_p(x)\right] \\ &= 0 + f(x) = f(x), \end{aligned}$$

where the first equality follows from (6), with $\alpha = \beta = 1$, and u, v equal to y_h and y_p, respectively.

To see that it is a general solution, let y be any solution of (1). Again using the linearity of L, we have

$$L\left[y - y_p\right] = L\left[y\right] - L\left[y_p\right] = f - f = 0,$$

so that the most general $y - y_p$ is a linear combination of a fundamental set of solutions of the homogeneous version of (1), namely y_h. Hence $y = y_h + y_p$ is a general solution of (1). ■

Thus, to solve the nonhomogeneous equation (1) we need to augment the homogeneous solution $y_h(x)$ by adding to it any particular solution $y_p(x)$.

Often, in applications, $f(x)$ is not a single term but a linear combination of terms: $f(x) = f_1(x) + \cdots + f_k(x)$. In the equation $L[y] = 5x^2 - 2\sin x + 6$, for instance, we can identify $f_1(x) = 5x^2$, $f_2 = -2\sin x$, and $f_3(x) = 6$.

THEOREM 3.7.2 *General Solution of* $L[y] = f_1 + \cdots + f_k$
If $y_h(x)$ is a general solution of $L[y] = 0$ on an interval I, and $y_{p1}(x), \dots, y_{pk}(x)$ are particular solutions of $L[y] = f_1, \dots, L[y] = f_k$ on I, respectively, then a general solution of $L[y] = f_1 + \cdots + f_k$ on I is

$$y(x) = y_h(x) + y_{p1}(x) + \cdots + y_{pk}(x). \tag{9}$$

Proof: That (9) satisfies (1) follows from (7), with all the α's equal to 1:

$$\begin{aligned} L\left[y_h + y_{p1} + \cdots + y_{pk}\right] &= L\left[y_h\right] + L\left[y_{p1}\right] + \cdots + L\left[y_{pk}\right] \\ &= 0 + f_1 + \cdots + f_k \\ &= f_1 + \cdots + f_k, \end{aligned} \tag{10}$$

as was to be verified.

To see that it is a *general* solution of (1), let y be any solution of (1). Then

$$\begin{aligned} L\left[y - y_{p1} - \cdots - y_{pk}\right] &= L\left[y\right] - L\left[y_{p1}\right] - \cdots - L\left[y_{pk}\right] \\ &= f - f_1 - \cdots - f_k = 0, \end{aligned}$$

so the most general $y - y_{p1} - \cdots - y_{pk}$ is a general solution y_h of the homogeneous version of (1). ■

This result is a **superposition principle**. It tells us that the response y_p to a superposition of inputs (the forcing functions $f_1, \dots, f_k$) is the superposition of their individual outputs $(y_{p1}, \dots, y_{pk})$.

The upshot is that to solve a nonhomogeneous equation we need to find both the homogeneous solution y_h and a particular solution y_p. Having already developed methods for determining y_h – at least for certain types of equations – we now need to present methods for determining particular solutions y_p, and that is the subject of Sections 3.7.2 and 3.7.3 that follow.

3.7.2. Undetermined coefficients. The method of undetermined coefficients is a procedure for determining a particular solution to the linear equation

$$\begin{aligned} L[y] &= f(x) \\ &= f_1(x) + \cdots + f_k(x), \end{aligned} \tag{11}$$

subject to two conditions:

(i) Besides being linear, L is of constant-coefficient type.

(ii) Repeated differentiation of each $f_j(x)$ term in (11) produces only a finite number of LI (linearly independent) terms.

To explain the latter condition, consider $f_j(x) = 2xe^{-x}$. The sequence consisting of this term and its successive derivatives is

$$2xe^{-x} \longrightarrow \left\{2xe^{-x},\ 2e^{-x} - 2xe^{-x},\ -4e^{-x} + 2xe^{-x},\ \dots\right\},$$

and we can see that this sequence contains only the two LI functions e^{-x} and xe^{-x}. Thus, $f_j(x) = 2xe^{-x}$ satisfies condition (ii).

As a second example, consider $f_j(x) = x^2$. This term generates the sequence

$$x^2 \longrightarrow \left\{x^2, 2x, 2, 0, 0, \dots\right\},$$

which contains only the three LI functions x^2, x, and 1. Thus, $f_j(x) = x^2$ satisfies condition (ii).

The term $f_j(x) = 1/x$, however, generates the sequence

$$1/x \longrightarrow \left\{1/x, -1/x^2, 2/x^3, -6/x^4, \dots\right\},$$

which contains an infinite number of LI terms ($1/x, 1/x^2, 1/x^3, \ldots$). Thus, $f_j(x) = 1/x$ does not satisfy condition (ii).

If the term $f_j(x)$ does satisfy condition (ii), then we will call the finite set of LI terms generated by it, through repeated differentiation, the **family generated by $f_j(x)$**. (That terminology is for convenience here and is not standard.) Thus, the family generated by $2xe^{-x}$ is comprised of e^{-x} and xe^{-x}, and the family generated by $3x^2$ is comprised of x^2, x, and 1.

Let us now illustrate the method of undetermined coefficients.

EXAMPLE 1. Consider the differential equation

$$y'''' - y'' = 3x^2 - \sin 2x. \tag{12}$$

First, we see that L is linear, with constant coefficients, so condition (i) is satisfied. Next, we identify $f_1(x), f_2(x)$, and their generated sequences as

$$f_1(x) = 3x^2 \longrightarrow \{3x^2, 6x, 6, 0, 0, \ldots\}, \tag{13a}$$
$$f_2(x) = -\sin 2x \longrightarrow \{-\sin 2x, -2\cos 2x, 4\sin 2x, \ldots\}. \tag{13b}$$

Thus, f_1 and f_2 do generate the finite families

$$f_1(x) = 3x^2 \longrightarrow \{x^2, x, 1\}, \tag{14a}$$
$$f_2(x) = -\sin 2x \longrightarrow \{\sin 2x, \cos 2x\}, \tag{14b}$$

so condition (ii) is satisfied.

To find a particular solution y_{p1} corresponding to f_1, tentatively seek it as a linear combination of the terms in (14a):

$$y_{p1}(x) = Ax^2 + Bx + C, \tag{15}$$

where A, B, C are the so-called **undetermined coefficients**. Next, we write down the homogeneous solution of (12),

$$y_h(x) = C_1 + C_2x + C_3e^x + C_4e^{-x}, \tag{16}$$

and check each term in y_{p1} [i.e., in (15)] for duplication with terms in y_h. Doing so, we find that the Bx and C terms in (15) duplicate (to within constant scale factors) the C_2x and C_1 terms, respectively, in (16). The method then calls for multiplying the entire family, involved in the duplication, by the lowest positive integer power of x needed to eliminate all such duplication. Thus, we revise (15) as

$$y_{p1}(x) = x\left(Ax^2 + Bx + C\right) = Ax^3 + Bx^2 + Cx, \tag{17}$$

but find that the Cx term in (17) is still "in violation" in that it duplicates the C_2x term in (16). Thus, try

$$y_{p1}(x) = x^2\left(Ax^2 + Bx + C\right) = Ax^4 + Bx^3 + Cx^2. \tag{18}$$

This time we are done, since all duplication has now been eliminated.

Next, we put the final revised form (18) into the equation $y'''' - y'' = 3x^2$ [i.e., $L[y] = f_1(x)$] and obtain

$$24A - 12Ax^2 - 6Bx - 2C = 3x^2. \tag{19}$$

Finally, equating coefficients of like terms gives

$$\begin{aligned} x^2: &\quad -12A = 3 \\ x: &\quad -6B = 0 \\ 1: &\quad 24A - 2C = 0, \end{aligned} \tag{20}$$

so that $A = -1/4$, $B = 0$, $C = -3$. Thus

$$y_{p1}(x) = -\frac{1}{4}x^4 - 3x^2. \tag{21}$$

Next, we need to find y_{p2} corresponding to f_2. To do so, we seek it as a linear combination of the terms in (14b):

$$y_{p2}(x) = D\sin 2x + E\cos 2x. \tag{22}$$

Checking each term in (22) for duplication with terms in y_h, we see that there is no such duplication. Thus, we accept the form (22), put it into the equation $y'''' - y'' = -\sin 2x$ [i.e., $L[y] = f_2(x)$], and obtain

$$20D\sin 2x + 20E\cos 2x = -\sin 2x. \tag{23}$$

Equating coefficients of like terms gives $20D = -1$, and $20E = 0$, so that $D = -1/20$ and $E = 0$. Thus,

$$y_{p2}(x) = -\frac{1}{20}\sin 2x. \tag{24}$$

Finally, a general solution of (12) is, according to Theorem 3.7.2,

$$y(x) = y_h(x) + y_p(x) = y_h(x) + y_{p1}(x) + y_{p2}(x),$$

namely,

$$y(x) = C_1 + C_2x + C_3e^x + C_4e^{-x} - \frac{1}{4}x^4 - 3x^2 - \frac{1}{20}\sin 2x. \tag{25}$$

COMMENT 1. We obtained (20) by "equating coefficients of like terms" in (19). That step amounted to using the concept of linear independence – namely, noting that $1, x, x^2$ are LI (on any given x interval) and then using Theorem 3.2.6. Alternatively, we could have rewritten (19) as

$$(24A - 2C)1 + (-6B)x + (-12A - 3)x^2 = 0 \tag{26}$$

and used the linear independence of $1, x, x^2$ to infer that the coefficient of each term must be zero, which step once again gives (20).

COMMENT 2. The key point in the analysis is that the system (20), consisting of three linear algebraic equations in the three unknowns A, B, C, is consistent. Similarly for the system $20D = -1$, $20E = 0$ governing D and E. The guarantee provided by the method of undetermined coefficients is that the resulting system of linear algebraic equations on the unknown coefficients will indeed be consistent, so that we can successfully solve for the undetermined coefficients. What would have happened if we had used the form (15) for $y_p(x)$ instead of (18) – that is, if we had not adhered to the prescribed procedure? We would have obtained, in place of (20), the equations

$$\begin{aligned} x^2: &\quad 0 = 3 \\ x: &\quad 0 = 0 \\ 1: &\quad -2A = 0, \end{aligned}$$

which are *in*consistent because the "equation" $0 = 3$ cannot be satisfied by any choice of A, B, C. That is, (15) would not have worked. ▮

Let us summarize.

STEPS IN THE METHOD OF UNDETERMINED COEFFICIENTS:

1. Verify that condition (i) is satisfied.
2. Identify the $f_j(x)$'s and verify that each one satisfies condition (ii).
3. Determine the finite family corresponding to each $f_j(x)$ [(14a,b) in Example 1].
4. Seek $y_{p1}(x)$, tentatively, as a linear combination of the terms in the family corresponding to $f_1(x)$ [(15) in Example 1].
5. Obtain the general solution $y_h(x)$ of the homogeneous equation [(16) in Example 1].
6. If such duplication is found, multiply the entire linear combination of terms by the lowest positive integer power of x necessary to remove all such duplication between those terms and the terms in y_h [(18) in Example 1].
7. Substitute the final version of the form assumed for $y_{p1}(x)$ into the left-hand side of the equation $L[y] = f_1$, and equate coefficients of like terms.
8. Solve the resulting system of linear algebraic equations for the undetermined coefficients. That step completes our determination of $y_{p1}(x)$.
9. Repeat steps 4–8 for $y_{p2}, \ldots, y_{pk}$.
10. Then the general solution of $L[y] = f_1 + \cdots + f_k$ is given, according to Theorem 3.7.2, by $y(x) = y_h(x) + y_p(x) = y_h(x) + y_{p1}(x) + \cdots + y_{pk}(x)$.

EXAMPLE 2. As a final example, consider

$$y'' - 9y = 4 + 5\sinh 3x, \tag{27}$$

which is indeed linear and of constant-coefficient type. Since

$$\begin{aligned} f_1(x) = 4 &\longrightarrow \{4, 0, 0, \ldots\}, \\ f_2(x) = 5\sinh 3x &\longrightarrow \{5\sinh 3x, 15\cosh 3x, 45\sinh 3x, \ldots\}, \end{aligned}$$

we see that these terms generate the finite families

$$\begin{aligned} f_1(x) = 4 &\longrightarrow \{1\}, \\ f_2(x) = 5\sinh 3x &\longrightarrow \{\sinh 3x, \cosh 3x\}, \end{aligned}$$

so we tentatively seek

$$y_{p1}(x) = A. \tag{28}$$

Since

$$y_h(x) = C_1 e^{3x} + C_2 e^{-3x}, \tag{29}$$

there is no duplication between the term in (28) and those in (29). Putting (28) into $y'' - 9y = 4$ gives $-9A = 4$, so $A = -4/9$. Thus,

$$y_{p1}(x) = -\frac{4}{9}. \tag{30}$$

Next, we tentatively seek

$$y_{p2}(x) = B\sinh 3x + C\cosh 3x. \tag{31}$$

At first glance, it appears that there is no duplication between any of the terms in (31) and those in (29). However, since the $\sinh 3x$ and $\cosh 3x$ are linear combinations of e^{3x} and e^{-3x}, we do indeed have duplication. Said differently, each of the $\sinh 3x$ and $\cosh 3x$ terms are solutions of the homogeneous equation. Thus, we need to multiply the right-hand side of (31) by x and revise y_p as

$$y_{p2}(x) = x\left(B\sinh 3x + C\cosh 3x\right). \tag{32}$$

Now that y_{p2} is in a satisfactory form we put that form into $y'' - 9y = 5\sinh 3x$ [i.e., $L[y] = f_2(x)$] and obtain the equation

$$\begin{aligned} &(3C + 3C)\sinh 3x + (3B + 3B)\cosh 3x \\ &+(9B - 9B)x\sinh 3x + (9C - 9C)x\cosh 3x = 5\sinh 3x. \end{aligned}$$

Equating coefficients of like terms gives $B = 5/6$ and $C = 0$, so

$$y_{p2}(x) = \frac{5}{6}x\sinh 3x. \tag{33}$$

It follows then, from Theorem 3.7.2, that a general solution of (27) is

$$y(x) = C_1 e^{3x} + C_2 e^{-3x} - \frac{4}{9} + \frac{5}{6}x\sinh 3x. \tag{34}$$

Naturally, one's final result can (and should) be checked by direct substitution into the original differential equation.

COMMENT. Suppose that in addition to the differential equation (27), initial conditions $y(0) = 0, y'(0) = 2$ are specified. Imposing these conditions on (34) gives $C_1 = 5/9, C_2 = -1/9$, and hence the particular solution

$$y(x) = \frac{5}{9}e^{3x} - \frac{1}{9}e^{-3x} - \frac{4}{9} + \frac{5}{6}x \sinh 3x. \tag{35}$$

Do not be concerned that we call (35) a particular solution even though each of the two exponential terms in (35) is a homogeneous solution because if we put (35) into the left-hand side of (27) it does give the right-hand side of (27); thus, it is a particular solution of (27). ▮

As a word of caution, suppose the differential equation is

$$y'' - 3y' + 2y = 2 \sinh x,$$

with homogeneous solution $C_1e^x + C_2e^{2x}$. Observe that $2 \sinh x = e^x - e^{-x}$ contains an e^x term, which corresponds to one of the homogeneous solutions. To bring this duplication into the light, we should re-express the differential equation as

$$y'' - 3y' + 2y = e^x - e^{-x}$$

before beginning the method of undetermined coefficients. Then, the particular solution due to $f_1(x) = e^x$ will be $y_{p1}(x) = Axe^x$ and the assumed particular solution due to $f_2(x) = e^{-x}$ will be $y_{p2}(x) = Be^{2x}$. We find that $A = -1$ and $B = -1/6$ so the general solution is

$$y(x) = C_1e^x + C_2e^{2x} - xe^x - \frac{1}{6}e^{-x}.$$

Closing this discussion of the method of undetermined coefficients, let us reconsider condition (ii), that repeated differentiation of each term in the forcing function must produce only a finite number of LI terms. How broad is the class of functions that satisfy that condition? If a forcing function f satisfies that condition, then it must be true that coefficients a_j exist, not all of them zero, such that

$$a_0f^{(N)} + a_1f^{(N-1)} + \cdots + a_{N-1}f' + a_Nf = 0 \tag{36}$$

over the x interval under consideration. From our discussion of the solution of such constant-coefficient equations we know that solutions f of (36) must be of the form $Cx^me^{(\alpha+i\beta)x}$, or a linear combination of such terms. Such functions are so common in applications that condition (ii) is not as restrictive as it may seem.

3.7.3. Variation of parameters. Although easy to apply, the method of undetermined coefficients is limited by the two conditions cited above – that L be of constant-coefficient type, and that repeated differentiation of each $f_j(x)$ forcing term produces only a finite number of LI terms.

The method of variation of parameters, due to Lagrange, is more powerful in that it is not subject to those restrictions. As with automobile engines, we can expect more power to come at a higher price and, as we shall see, Lagrange's method is indeed the more difficult to apply.

In fact, we have already presented the method, in Section 2.2, for the general linear first-order equation

$$y' + p(x)y = q(x), \tag{37}$$

and we urge you to review that discussion. The idea was to seek a particular solution y_p by varying the parameter A (i.e., the constant of integration) in the homogeneous solution

$$y_h(x) = Ae^{-\int p(x)\,dx}.$$

Thus, we sought

$$y_p(x) = A(x)e^{-\int p(x)\,dx},$$

put that form into (37) and solved for $A(x)$.

Likewise, if an nth-order linear differential equation $L[y] = f$ has a homogeneous solution

$$y_h(x) = C_1y_1(x) + \cdots + C_ny_n(x), \tag{38}$$

then according to the method of variation of parameters we seek a particular solution in the form

$$y_p(x) = C_1(x)y_1(x) + \cdots + C_n(x)y_n(x); \tag{39}$$

that is, we "vary the parameters" (constants of integration) $C_1, \ldots, C_n$ in (38).

Let us carry out the procedure for the linear second-order equation

$$L[y] = y'' + p_1(x)y' + p_2(x)y = f(x), \tag{40}$$

where the coefficients $p_1(x)$ and $p_2(x)$ are assumed to be continuous on the x interval of interest, say I. We suppose that

$$y_h(x) = C_1y_1(x) + C_2y_2(x) \tag{41}$$

is a known general solution of the homogeneous equation on I, and seek

$$y_p(x) = C_1(x)y_1(x) + C_2(x)y_2(x). \tag{42}$$

Needing y_p' and y_p'', to substitute into (40), we differentiate (42):

$$y_p' = C_1y_1' + C_2y_2' + C_1'y_1 + C_2'y_2. \tag{43}$$

Looking ahead, y_p'' will include $C_1, C_2, C_1', C_2', C_1'', C_2''$ terms, so that (40) will become a nonhomogeneous second-order differential equation in C_1 and C_2, which can hardly be expected to be simpler than the original equation (40)! However, it will be only one equation in the two unknowns C_1, C_2, so we are free to impose another condition on C_1, C_2 to complete, and simplify, the system.

An especially convenient condition to impose will be

$$C_1'y_1 + C_2'y_2 = 0, \tag{44}$$

for this condition will knock out the C_1', C_2' terms in (43), so that y_p'' will contain only *first*-order derivatives of C_1 and C_2. Then (43) reduces to $y_p' = C_1y_1' + C_2y_2'$, so

$$y_p'' = C_1y_1'' + C_2y_2'' + C_1'y_1' + C_2'y_2', \tag{45}$$

and (40) becomes

$$C_1\left(y_1'' + p_1y_1' + p_2y_1\right) + C_2\left(y_2'' + p_1y_2' + p_2y_2\right) + C_1'y_1' + C_2'y_2' = f. \tag{46}$$

The two parenthetic groups vanish by virtue of y_1 and y_2 being solutions of the homogeneous equation $L[y] = 0$, so (46) simplifies to $C_1'y_1' + C_2'y_2' = f$. That result, together with (44), gives us the equations

$$\begin{aligned} y_1C_1' + y_2C_2' &= 0, \\ y_1'C_1' + y_2'C_2' &= f \end{aligned} \tag{47}$$

on C_1', C_2'. The latter will be solvable for C_1', C_2', uniquely, if the determinant of the coefficients does not vanish on I. In fact, we recognize that determinant as the Wronskian of y_1 and y_2,

$$W[y_1, y_2](x) = \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix}, \tag{48}$$

and the latter is necessarily nonzero on I by Theorem 3.2.3 because y_1 and y_2 are LI solutions of $L[y] = 0$.

Solving (47) by Cramer's rule gives

$$C_1'(x) = \frac{\begin{vmatrix} 0 & y_2 \\ f & y_2' \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}} = \frac{W_1(x)}{W(x)}, \quad C_2'(x) = \frac{\begin{vmatrix} y_1 & 0 \\ y_1' & f \end{vmatrix}}{\begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}} = \frac{W_2(x)}{W(x)},$$

where W_1, W_2 simply denote the determinants in the numerators. Integrating these equations and putting the results into (42) gives

$$\boxed{y_p(x) = \left[\int^x \frac{W_1(\xi)}{W(\xi)}\,d\xi\right] y_1(x) + \left[\int^x \frac{W_2(\xi)}{W(\xi)}\,d\xi\right] y_2(x),} \tag{49}$$

or, more compactly,

$$y_p(x) = \int^x \frac{W_1(\xi)y_1(x) + W_2(\xi)y_2(x)}{W(\xi)}\,d\xi. \tag{50}$$

EXAMPLE 3. To solve

$$y'' - 4y = 8e^{2x}, \tag{51}$$

we note the general solution

$$y_h(x) = C_1e^{2x} + C_2e^{-2x} \tag{52}$$

of the homogeneous equation, so that we may take $y_1(x) = e^{2x}, y_2(x) = e^{-2x}$. Then $W(x) = y_1y_2' - y_1'y_2 = -4, W_1(x) = -f(x)y_2(x) = -8e^{2x}e^{-2x} = -8$, and $W_2(x) = f(x)y_1(x) = 8e^{2x}e^{2x} = 2e^{4x}$, so (49) gives

$$\begin{aligned} y_p(x) &= \left(\int^x 2\,d\xi\right)e^{2x} + \left(\int^x -2e^{4\xi}\,d\xi\right)e^{-2x} \\ &= (2x + A)e^{2x} + \left(-\frac{e^{4x}}{2} + B\right)e^{-2x} = 2xe^{2x} - \frac{e^{2x}}{2} + Ae^{2x} + Be^{-2x}, \end{aligned} \tag{53}$$

where A, B are the arbitrary constants of integration. We can omit the A, B terms in (53) because they give terms (Ae^{2x} and Be^{-2x}) that merely duplicate those already present in the homogeneous solution y_h. That will always be the case: we can omit the constants of integration in the two integrals in (49). In the present example we can even drop the $-e^{2x}/2$ term in the right side of (53) since it too is a homogeneous solution [and can be absorbed in the C_1e^{2x} term in (55)]. Thus, we write

$$y_p(x) = 2xe^{2x}. \tag{54}$$

Finally,

$$y(x) = y_h(x) + y_p(x) = C_1e^{2x} + C_2e^{-2x} + 2xe^{2x} \tag{55}$$

gives a general solution of (51). ▮

3.7.4. Variation of parameters for higher-order equations. (Optional) For higher-order equations the idea is essentially the same. For the third-order equation

$$L[y] = y''' + p_1(x)y'' + p_2(x)y' + p_3(x)y = f(x), \tag{56}$$

for instance, if

$$y_h(x) = C_1y_1(x) + C_2y_2(x) + C_3y_3(x) \tag{57}$$

is known, then we seek

$$y_p(x) = C_1(x)y_1(x) + C_2(x)y_2(x) + C_3(x)y_3(x). \tag{58}$$

Looking ahead, when we put (58) into (56) we will have one equation in the three unknown functions C_1, C_2, C_3, so we can impose two additional conditions on C_1, C_2, C_3 to complete, and simplify, the system. Proceeding, differentiation of (58) gives

$$y_p'(x) = C_1 y_1' + C_2 y_2' + C_3 y_3' + C_1' y_1 + C_2' y_2 + C_3' y_3, \tag{59}$$

so we set

$$C_1' y_1 + C_2' y_2 + C_3' y_3 = 0 \tag{60}$$

to suppress higher-order derivatives of C_1, C_2, C_3. Then (59) reduces to

$$y_p' = C_1 y_1' + C_2 y_2' + C_3 y_3', \tag{61}$$

and another differentiation gives

$$y_p'' = C_1 y_1'' + C_2 y_2'' + C_3 y_3'' + C_1' y_1' + C_2' y_2' + C_3' y_3'. \tag{62}$$

Again, to suppress higher-order derivatives of C_1, C_2, C_3, set

$$C_1' y_1' + C_2' y_2' + C_3' y_3' = 0. \tag{63}$$

Then (62) reduces to

$$y_p'' = C_1 y_1'' + C_2 y_2'' + C_3 y_3'', \tag{64}$$

so

$$y_p''' = C_1 y_1''' + C_2 y_2''' + C_3 y_3''' + C_1' y_1'' + C_2' y_2'' + C_3' y_3''. \tag{65}$$

Finally, putting (65), (64), (61), and (58) into (56) gives

$$\begin{aligned} &C_1 \left(y_1''' + p_1 y_1'' + p_2 y_1' + p_3 y_1\right) + C_2 \left(y_2''' + p_1 y_2'' + p_2 y_2' + p_3 y_2\right) \\ &\quad + C_1 \left(y_3''' + p_1 y_3'' + p_2 y_3' + p_3 y_3\right) + C_1' y_1'' + C_2' y_2'' + C_3' y_3'' = f, \end{aligned} \tag{66}$$

or

$$C_1' y_1'' + C_2' y_2'' + C_3' y_3'' = f \tag{67}$$

since each of the three parenthetic groups in (66) vanishes because y_1, y_2, y_3 are homogeneous solutions.

Equations (60), (63), and (67) are three linear algebraic equations in C_1', C_2', C_3':

$$\begin{aligned} y_1 C_1' + y_2 C_2' + y_3 C_3' &= 0, \\ y_1' C_1' + y_2' C_2' + y_3' C_3' &= 0, \\ y_1'' C_1' + y_2'' C_2' + y_3'' C_3' &= f. \end{aligned} \tag{68}$$

That system is uniquely solvable for C_1', C_2', C_3' because the determinant of the coefficients is the Wronskian determinant of y_1, y_2, y_3,

$$W[y_1, y_2, y_3](x) = \begin{vmatrix} y_1 & y_2 & y_3 \\ y_1' & y_2' & y_3' \\ y_1'' & y_2'' & y_3'' \end{vmatrix}, \tag{69}$$

which is nonzero on the interval I because y_1, y_2, y_3 are LI on I by the assumption that (57) is a general solution of the homogeneous equation $L[y] = 0$. Solving (68) by means of Cramer's rule gives

$$C_1' = \frac{\begin{vmatrix} 0 & y_2 & y_3 \\ 0 & y_2' & y_3' \\ f & y_2'' & y_3'' \end{vmatrix}}{W(x)} = \frac{W_1(x)}{W(x)}, \qquad C_2' = \frac{\begin{vmatrix} y_1 & 0 & y_3 \\ y_1' & 0 & y_3' \\ y_1'' & f & y_3'' \end{vmatrix}}{W(x)} = \frac{W_2(x)}{W(x)},$$
$$C_3' = \frac{\begin{vmatrix} y_1 & y_2 & 0 \\ y_1' & y_2' & 0 \\ y_1'' & y_2'' & f \end{vmatrix}}{W(x)} = \frac{W_3(x)}{W(x)}. \tag{70}$$

Finally, integrating these equations and putting the results into (58) gives the particular solution

$$\boxed{\begin{aligned} y_p(x) = {} & \left[\int^x \frac{W_1(\xi)}{W(\xi)}\,d\xi\right] y_1(x) + \left[\int^x \frac{W_2(\xi)}{W(\xi)}\,d\xi\right] y_2(x) \\ & + \left[\int^x \frac{W_3(\xi)}{W(\xi)}\,d\xi\right] y_3(x). \end{aligned}} \tag{71}$$

EXAMPLE 4. Consider the nonhomogeneous Cauchy–Euler equation

$$x^3y''' + x^2y'' - 2xy' + 2y = \frac{2}{x}. \qquad (0 < x < \infty) \tag{72}$$

Observe that we cannot use the method of undetermined coefficients in this case because the differential operator is not of constant-coefficient type, and also because the forcing function does not generate only a finite number of linearly-independent derivatives.

To use (71), we need to know y_1, y_2, y_3, and f. It is readily found that the homogeneous solution is

$$y_h(x) = C_1\frac{1}{x} + C_2x + C_3x^2, \tag{73}$$

so we can take $y_1 = 1/x$, $y_2 = x$, $y_3 = x^2$. But be careful: $f(x)$ is not $2/x$ because (72) is not yet in the form of (56). That is, (72) must be divided by x^3 so that the coefficient of y''' becomes 1, as in (56). Doing so, it follows that $f(x) = 2/x^4$.

We can now evaluate the determinants needed in (71):

$$W(x) = \begin{vmatrix} x^{-1} & x & x^2 \\ -x^{-2} & 1 & 2x \\ 2x^{-3} & 0 & 2 \end{vmatrix} = \frac{6}{x}, \qquad W_1(x) = \begin{vmatrix} 0 & x & x^2 \\ 0 & 1 & 2x \\ 2x^{-4} & 0 & 2 \end{vmatrix} = \frac{2}{x^2},$$

$$W_2(x) = \begin{vmatrix} x^{-1} & 0 & x^2 \\ -x^{-2} & 0 & 2x \\ 2x^{-3} & 2x^{-4} & 2 \end{vmatrix} = -\frac{2}{x^4}, \; W_3(x) = \begin{vmatrix} x^{-1} & x & 0 \\ -x^{-2} & 1 & 0 \\ 2x^{-3} & 0 & 2x^{-4} \end{vmatrix} = \frac{4}{x^5}, \tag{74}$$

so (71) gives

$$\begin{aligned} y_p(x) &= \left(\int^x \frac{1}{3\xi}\, d\xi\right)\frac{1}{x} + \left(\int^x -\frac{1}{3\xi^3}\, d\xi\right)x + \left(\int^x \frac{2}{3\xi^4}\, d\xi\right)x^2 \\ &= \frac{\ln x}{3x} - \frac{1}{18x}. \end{aligned} \tag{75}$$

The $-1/(18x)$ term can be dropped because it is a homogeneous solution, so

$$y_p(x) = \frac{1}{3}\frac{\ln x}{x}, \tag{76}$$

as can be verified by direct substitution into (56). ▮

Generalization of the method to the nth-order equation

$$y^{(n)} + p_1(x)y^{(n-1)} + \cdots + p_{n-1}(x)y' + p_n(x)y = f(x) \tag{77}$$

is straightforward and the result is

$$\boxed{y_p(x) = \left[\int^x \frac{W_1(\xi)}{W(\xi)}\, d\xi\right] y_1(x) + \cdots + \left[\int^x \frac{W_n(\xi)}{W(\xi)}\, d\xi\right] y_n(x),} \tag{78}$$

where the y_j's are n LI homogeneous solutions, W is the Wronskian of $y_1, \ldots, y_n$, and W_j is identical to W, but with the jth column replaced by a column of zeros – except for the bottom element, which is f.

Closure. In this section we have discussed the nonhomogeneous equation $L[y] = f$, where L is an nth-order linear differential operator.

In Section 3.7.1 we provided the theoretical framework, which was based upon the linearity of L. Specifically, Theorem 3.7.1 showed that the general solution of $L[y] = f$ can be formed as the sum of a homogeneous solution $y_h(x)$, and a particular solution $y_p(x)$; $y_h(x)$ is a general solution of $L[y] = 0$, so it contains the n constants of integration, and $y_p(x)$ is any particular solution of the full equation $L[y] = f$. Further, we showed that if f is broken down as $f = f_1 + \cdots + f_k$, then y_p is the sum of particular solutions $y_{p1}, \dots, y_{pk}$ corresponding to $f_1, \dots, f_k$, respectively.

Having studied homogeneous equations earlier, we considered y_h as known and focused our attention, in Sections 3.7.2 and 3.7.3, on methods for finding particular solutions y_p: undetermined coefficients and variation of parameters. Of these, the former is the easier to implement, but it is limited to cases where L is of constant-coefficient type and where each $f_j(x)$ has only a finite number of LI derivatives. The latter is harder to apply since it requires integrations, but is more powerful since it works even if L has nonconstant coefficients, and for any functions $f_j(x)$.

EXERCISES 3.7

1. Show whether or not the given forcing function satisfies condition (ii), below equation (11). If so, give a finite family of LI functions generated by it.

(a) $x^2 \cos x$ (b) $\cos x \sinh 2x$
(c) $\ln x$ (d) $x^2 \ln x$
(e) $\sin x / x$ (f) e^{x^2}
(g) e^{9x} (h) $(x-1)/(x+2)$
(i) $\tan x$ (j) $e^x \cos 3x$
(k) $x^3 e^{-x} \sinh x$ (l) $\cos x \cos 2x$
(m) $\sin x \sin 2x \sin 3x$ (n) $e^x/(x+1)$

2. Obtain a general solution using the method of undetermined coefficients.

(a) $y' - 3y = xe^{2x} + 6$
(b) $y' + y = x^4 + 2x$
(c) $y' + 2y = 3e^{2x} + 4\sin x$
(d) $y' - 3y = xe^{3x} + 4$
(e) $y' + y = 5 - e^{-x}$
(f) $y' - y = x^2 e^x$
(g) $y'' - y' = 5 \sin 2x$
(h) $y'' + y' = 4xe^x + 3\sin x$
(i) $y'' + y = 3\sin 2x - 5 + 2x^2$
(j) $y'' + y' - 2y = x^3 - e^{-x}$
(k) $y'' + y = 6\cos x + 2$
(l) $y'' + 2y' = x^2 + 4e^{2x}$
(m) $y'' - 2y' + y = x^2 e^x$
(n) $y'' - 4y = 5(\cosh 2x - x)$
(o) $y'' - y' = 2xe^x$
(p) $y''' - y' = 25\cos 2x$
(q) $y''' - y'' = 6x + 2\cosh x$
(r) $y'''' + y'' - 2y = 3x^2 - 1$
(s) $y'''' - y = 5(x + \cos x)$

3. (a)–(s) Use computer software to solve the corresponding problem in Exercise 2.

4. Obtain a general solution using the method of variation of parameters.

(a) $y' + 2y = 4e^{2x}$
(b) $y' - y = xe^x + 1$
(c) $xy' - y = x^3$
(d) $xy' + y = 1/x \quad (x > 0)$
(e) $x^3 y' + x^2 y = 1 \quad (x > 0)$
(f) $y'' - y = 8x$
(g) $y'' - y = 8e^x$
(h) $y'' - 2y' + y = 6x^2$
(i) $y'' - 2y' + y = 2e^x$
(j) $y'' + y = 4\sin x$
(k) $y'' + 4y' + 4y = 2e^{-2x}$
(l) $6y'' - 5y' + y = x^2$
(m) $x^3 y'' + x^2 y' - 4xy = 1 \quad (x > 0)$
(n) $x^2 y'' - xy' - 3y = 4x \quad (x < 0)$
(o) $y''' + y'' - y' - y = x$
(p) $y''' - 6y'' + 11y' - 6y = e^{4x}$

5. (a)–(p) Use computer software to solve the corresponding problem in Exercise 4.

6. In the method of variation of parameters we used indefinite integrals [in (49) and (78)]. However, we could use definite integrals in those formulas, instead, if we choose. Specifically, show that, in place of (49),

$$y_p(x) = \left[\int_{\alpha_1}^{x} \frac{W_1(\xi)}{W(\xi)}\, d\xi\right] y_1(x) + \left[\int_{\alpha_2}^{x} \frac{W_2(\xi)}{W(\xi)}\, d\xi\right] y_2(x)$$

is also correct, for any choice of the constants α_1, α_2 (although normally one would choose α_1 and α_2 to be the same).

7. We chose (44) as "an especially convenient" condition to impose on C_1 and C_2. The condition

$$C_1'y_1 + C_2'y_2 = 6,$$

say, is not as simple inasmuch as 6 is not as simple as 0, but would it work? If so, use it in place of (44) and derive the final result for y_p, in place of (49). If not, explain why it would not work.

3.8 Application to Harmonic Oscillator: Forced Oscillation

The free oscillation of the harmonic oscillator (Fig. 1) was studied in Section 3.5. Now that we know how to find particular solutions, we can return to the harmonic oscillator and consider the case of forced oscillations, governed by the second-order, linear, constant-coefficient, nonhomogeneous equation

$$mx'' + cx' + kx = f(t). \tag{1}$$

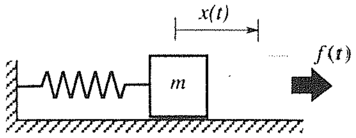


Figure 1. Mechanical oscillator.

In particular, we consider the important case where the forcing function is harmonic,

$$f(t) = F_0 \cos \Omega t. \tag{2}$$

3.8.1. Undamped case. To begin, consider the undamped case ($c = 0$),

$$mx'' + kx = F_0 \cos \Omega t. \tag{3}$$

The homogeneous solution of (3) is

$$x_h(t) = A \cos \omega t + B \sin \omega t, \tag{4}$$

where $\omega = \sqrt{k/m}$ is the natural frequency (i.e., the frequency of the free oscillation), and the forcing function $F_0 \cos \Omega t$ generates the family $\{\cos \Omega t, \sin \Omega t\}$. Thus, to find a particular solution of (3) by the method of undetermined coefficients, seek

$$x_p(t) = C \cos \Omega t + D \sin \Omega t. \tag{5}$$

Two cases present themselves. In the generic case, the driving frequency Ω is different from the natural frequency ω, so the terms in (5) do not duplicate any of those in (4) and we can accept (5) without modification. In the exceptional, or "singular," case where Ω is equal to ω, the terms in (5) repeat those in (4), so we need to modify (5) by multiplying the right side of (5) by t. For reasons that will become clear below, these cases are known as *nonresonance* and *resonance*, respectively.

Nonresonant oscillation. Putting (5) into the left side of (3) gives

$$\left(\omega^2 - \Omega^2\right) C \cos \Omega t + \left(\omega^2 - \Omega^2\right) D \sin \Omega t = \frac{F_0}{m} \cos \Omega t. \tag{6}$$

Since $\Omega \neq \omega$ by assumption, it follows from (6), by equating the coefficients of $\cos \Omega t$ and $\sin \Omega t$ on the left and right sides, that $C = (F_0/m)/(\omega^2 - \Omega^2)$ and $D = 0$. Thus

$$x_p(t) = \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t, \tag{7}$$

so a general solution of (3) is

$$\begin{aligned} x(t) &= x_h(t) + x_p(t) \\ &= A \cos \omega t + B \sin \omega t + \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t. \end{aligned} \tag{8}$$

In a sense we are done, and if we wish to impose any prescribed initial conditions $x(0)$ and $x'(0)$, then we could use those conditions to evaluate the constants A and B in (8). Then, for any desired numerical values of m, k, F_0, and Ω we could plot $x(t)$ versus t and see what the solution looks like. However, in science and engineering one is interested not only in obtaining answers, but also in understanding phenomena, so the question is: How can we extract, from (8), an understanding of the phenomenon? To answer that question, let us first rewrite (8) in the equivalent form

$$x(t) = E \sin(\omega t + \phi) + \frac{F_0/m}{\omega^2 - \Omega^2} \cos \Omega t \tag{9}$$

since then we can see it more clearly as a superposition of two harmonic solutions, of different amplitude, frequency, and phase.

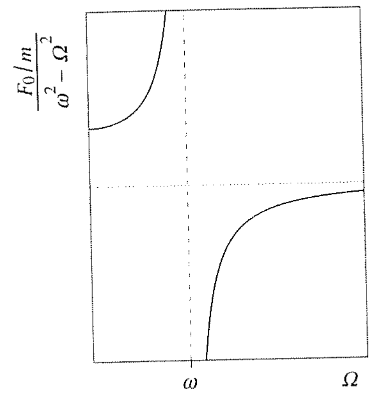


Figure 2. Magnitude of response (undamped case).

The homogeneous solution $E \sin(\omega t + \phi)$ in (9), the "free vibration," was already discussed in Section 3.5. [Alternative to $E \sin(\omega t + \phi)$, we could use the form $E \cos(\omega t + \phi)$, whichever one prefers; it doesn't matter.] Thus, consider the particular solution, or "forced response," given by (7) and the last term in (9). It is natural to regard m and k (and hence ω) as fixed, and F_0 and Ω as controllable quantities or parameters. That the response (7) is merely proportional to F_0 is no surprise, for it follows from the linearity of the differential operator in (3). We also see, from (7), that the response is at the same frequency as the forcing function, Ω. More interesting is the variation of the amplitude $(F_0/m)/(\omega^2 - \Omega^2)$ with Ω, which is sketched in Fig. 2. The change in sign, as Ω increases through ω, is awkward since it prevents us from interpreting the plotted quantity as a pure magnitude. Thus, let us re-express (7) in the equivalent form

$$x_p(t) = \frac{F_0/m}{|\omega^2 - \Omega^2|} \cos(\Omega t + \Phi), \tag{10}$$

where the phase angle Φ is 0 for $\Omega < \omega$ and π for $\Omega > \omega$ [since $\cos(\Omega t + \pi) =$

$-\cos\Omega t$ gives the desired sign change for $\Omega > \omega$]. The resulting amplitude- and phase-response curves are shown in Fig. 3. From Fig. 3a, observe that as the driving frequency approaches the natural frequency the amplitude tends to infinity! [Of course, we must remember that right *at* $\Omega = \omega$ our particular solution is invalid since (6) is then $(0)\cos\Omega t + (0)\sin\Omega t = (F_0/m)\cos\Omega t$, which cannot be satisfied.] Further, as $\Omega \to \infty$ the amplitude tends to zero. Finally, we see from Fig. 3b that the response is in-phase ($\Phi = 0$) with the forcing function for $\Omega < \omega$, but for all $\Omega > \omega$ it is 180° out-of-phase. This discontinuous jump is striking since only an infinitesimal change in Ω (from just below ω to just above it) produces a discontinuous change in the response.

Also of considerable interest phenomenologically is the possibility of what is known as *beats*, but we will postpone that dicussion until we have had a look at the special case of resonance.

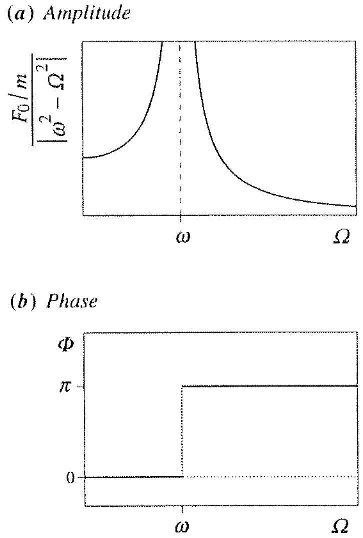


Figure 3. Amplitude- and phase-response curves (undamped case).

Resonant oscillation. For the special case where $\Omega = \omega$ (that is, where we force the system precisely at its natural frequency), the terms in (5) duplicate those in (4) so, according to the method of undetermined coefficients, we need to revise x_p as

$$x_p(t) = t\,(C\cos\omega t + D\sin\omega t)\,. \tag{11}$$

Since the duplication has thereby been removed, we accept (11). Putting that form into (3), we find that $C = 0$ and $D = F_0/(2m\omega)$, so

$$x_p(t) = \frac{F_0}{2m\omega} t \sin\omega t, \tag{12}$$

which is shown in Fig. 4. In this special case the response is not a harmonic oscillation but a harmonic function times t, which factor causes the magnitude to tend to infinity as $t \to \infty$. This result is known as **resonance**. Of course, the magnitude does not grow unboundedly in a real application since the mathematical model of the system (the governing differential equation) will become inaccurate for sufficiently large amplitudes, parts will break, and so on.

Resonance is sometimes welcome and sometimes unwelcome. That is, sometimes we wish to amplify a given input, and can do so by "tuning" the system to be at or near resonance, as when we tune a radio circuit to a desired broadcast frequency. And other times we wish to suppress inputs, as a well designed automobile suspension suppresses, rather than amplifies, the inputs from a bumpy road.

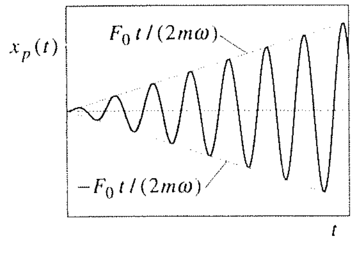


Figure 4. Resonant oscillation.

Beats. Isn't it striking that the response $x(t)$ is the sum of two harmonics [given by (5)] for all $\Omega \neq \omega$, yet it is of the different form (12) for the single case $\Omega = \omega$? One might wonder whether the resonant case is really of any importance at all since one can never get Ω to exactly equal ω. It is therefore of interest to look at the solution $x(t)$ as Ω *approaches* ω. To do so, let us use the simple initial conditions $x(0) = 0$ and $x'(0) = 0$, for definiteness, in which case we can evaluate A and B in (8), and obtain

$$x(t) = -\frac{F_0/m}{\omega^2 - \Omega^2}(\cos\omega t - \cos\Omega t)\,, \tag{13}$$

or, recalling the trigonometric identity $\cos A - \cos B = 2 \sin \dfrac{B+A}{2} \sin \dfrac{B-A}{2}$,

$$x(t) = \frac{2F_0/m}{\omega^2 - \Omega^2} \sin \left(\frac{\omega + \Omega}{2} \right) t \ \sin \left(\frac{\omega - \Omega}{2} \right) t. \tag{14}$$

Now, suppose that Ω is close to (but not equal to) the natural frequency ω. Then the frequency of the second sinusoid in (14) is very small compared to that of the first, so the $\sin \left(\frac{\omega - \Omega}{2} \right) t$ factor amounts, essentially, to a slow "amplitude modulation" of the relatively high frequency $\sin \left(\frac{\omega + \Omega}{2} \right) t$ factor. This phenomenon is known as **beats**, and is seen in Fig. 5, where we have plotted the solution (14) for four representative cases: in Fig. 5a Ω is not close to ω, and there is no discernible beat phenomenon, but as Ω is increased the beat phenomenon becomes well established, as seen in Fig. 5b, 5c, and 5d. [We have shown the "envelope" $\sin \left(\frac{\omega - \Omega}{2} \right) t$ as dotted.]

We can now see that the resonance phenomenon at $\Omega = \omega$ is not an isolated behavior but is a limiting case as $\Omega \to \omega$. That is, resonance (Fig. 4) is actually a limit of the sequence shown in Fig. 5, as $\Omega \to \omega$. Rather than depend only on these suggestive graphical results, we can proceed analytically as well. Specifically, we can take the limit of the response (13) as $\Omega \to \omega$ and, with the help of l'Hôpital's rule, we do obtain (12)!

With our foregoing discussion of the undamped forced harmonic oscillator in mind, *we cannot overstate that we are by no means dealing only with the solving of equations but with the phenomena thereby being described. To understand phenomena, we normally need to do several things: we do need to solve the equations that model the phenomena (analytically or, if that is too hard, numerically), but we also need to study, interpret, and understand the results.* Such study normally includes the generation of suitably chosen graphical displays (such as our Fig. 2, 3, and 4), the isolation of special cases [such as our initial consideration of the case where there is no damping; $c = 0$ in (1)], and perhaps the examination of various limiting cases (such as the limit $\Omega \to \omega$ in the present example). Emphasis in this book is on the mathematics, with the detailed study of the relevant physics left for applications courses such as Fluid Mechanics, Electromagnetic Field Theory, and so on, but we will occasionally try to show not only the connections between the mathematics and the physics but also the process whereby we determine those connections.

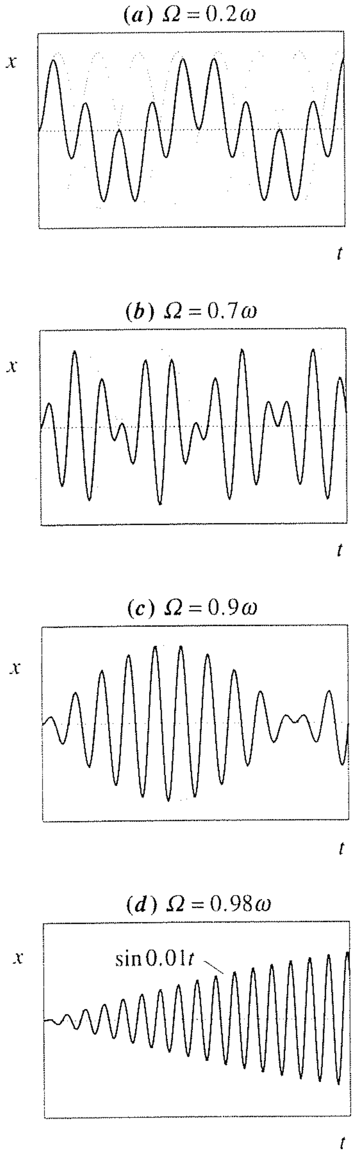


Figure 5. Beats, and approach to resonance.

3.8.2. Damped case. We now reconsider the harmonically driven oscillator, this time with a cx' damping term included ($c > 0$):

$$mx'' + cx' + kx = F_0 \cos \Omega t. \tag{15}$$

Recall from Section 3.5 that the homogeneous solution is

$$x_h(t) = \begin{cases} e^{-\frac{c}{2m}t}\left[A\cos\sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\,t + B\sin\sqrt{\omega^2 - \left(\frac{c}{2m}\right)^2}\,t\right] \\ e^{-\frac{c}{2m}t}(A + Bt) \\ e^{-\frac{c}{2m}t}\left[A\cosh\sqrt{\left(\frac{c}{2m}\right)^2 + \omega^2}\,t + B\sinh\sqrt{\left(\frac{c}{2m}\right)^2 + \omega^2}\,t\right] \end{cases} \tag{16}$$

for the underdamped ($c < c_{cr}$), critically damped ($c = c_{cr}$), and overdamped ($c > c_{cr}$) cases, respectively, and where $\omega = \sqrt{k/m}$ and $c_{cr} = 2\sqrt{mk}$.

This time, when we write

$$x_p(t) = C\cos\Omega t + D\sin\Omega t, \tag{17}$$

according to the method of undetermined coefficients, there is no duplication between terms in (17) and (16), even if $\Omega = \omega$, because of the $\exp(-ct/2m)$ factors in (16), so we can accept (17) without modification. Putting (17) into (15) and equating coefficients of the $\cos\Omega t$ terms on both sides of the equation, and similarly for the coefficients of the $\sin\Omega t$ terms, enables us to solve for C and D. The result (Exercise 3a) is that

$$\begin{aligned} x_p(t) = {} & \frac{(F_0/m)\left(\omega^2 - \Omega^2\right)}{\left(\omega^2 - \Omega^2\right)^2 + (c\,\Omega/m)^2}\cos\Omega t \\ & + \frac{F_0 c\,\Omega/m^2}{\left(\omega^2 - \Omega^2\right)^2 + (c\,\Omega/m)^2}\sin\Omega t, \end{aligned} \tag{18}$$

or (Exercise 3b), equivalently,

$$x_p(t) = E\cos(\Omega t + \Phi), \tag{19a}$$

where the amplitude E and phase Φ are

$$E = \frac{F_0/m}{\sqrt{\left(\omega^2 - \Omega^2\right)^2 + (c\,\Omega/m)^2}}, \tag{19b}$$

$$\Phi = \tan^{-1}\frac{c\,\Omega/m}{\Omega^2 - \omega^2}, \tag{19c}$$

with the $\tan^{-1}$ understood to lie between 0 and π.

As for the undamped case, we have great interest in the amplitude- and frequency-response curves, the graphs of the amplitude E, and the phase Φ with respect to the driving frequency Ω. The former is given in Fig. 6 for various values of the damping coefficient c, and the latter is left for the exercises.

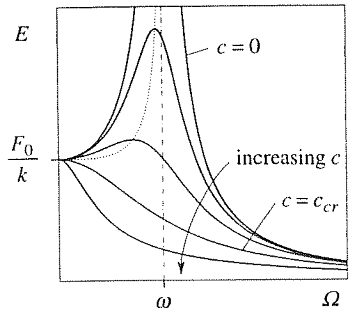


Figure 6. Amplitude response curves.

From Fig. 6 we see that true resonance is possible only in the case of no damping ($c = 0$), which case is an idealization since in reality there is inevitably some damping present. Analytically, we see the same thing: (19a) shows that the amplitude E can become infinite only if $c = 0$, and that occurs only for $\Omega = \omega$. However, for $c > 0$ there is still a peaking of the amplitude, even if that peak is now finite, at a driving frequency Ω which diminishes from ω as c increases, and which is 0 for all $c \geq c_{cr}$. Further, the peak magnitude (located by the dotted curve) diminishes from ∞ to F_0/k as c is increased from 0 to c_{cr}, and remains F_0/k for all $c > c_{cr}$.

What is the significance of the F_0/k value? For $\Omega = 0$ the differential equation becomes $mx'' + cx' + kx = F_0$, and the method of undetermined coefficients gives $x_p(t) = \text{constant} = F_0/k$, which is merely the static deflection of the mass under the steady force F_0.

Even if true resonance is possible only for the undamped case ($c = 0$), the term resonance is often used to refer to the dramatic peaking of the amplitude response curves if c is not too large.

The general solution, of course, is the sum

$$\begin{aligned} x(t) &= x_h(t) + x_p(t) \\ &= x_h(t) + E\cos(\Omega t + \Phi), \end{aligned} \tag{20}$$

where E and Φ are given by (19b,c) and $x_h(t)$ is given by the suitable right-hand side of (16), according to whether the system is underdamped, critically damped, or overdamped. If we impose initial conditions $x(0)$ and $x'(0)$ on (20), then we can solve for the integration constants A and B within $x_h(t)$.

Notice carefully that the $x_h(t)$ part of the solution inevitably tends to zero as $t \to \infty$ because of the $\exp(-ct/2m)$ factor, no matter how small c is, as long as $c > 0$. Thus, we call $x_h(t)$ in (20) the **transient** part of the solution and we call $x_p(t)$ the **steady-state** part since $x(t) \to E\cos(\Omega t + \Phi)$ as $t \to \infty$. The transient part depends upon the initial conditions, whereas the steady-state part does not. A representative underdamped case is shown in Fig. 7, where we see the approach to the steady-state oscillation $x_p(t)$.

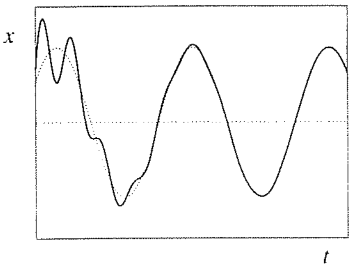


Figure 7. A representative response $x(t)$ (solid); approach to the steady-state oscillation $x_p(t)$ (dotted).

Closure. In this section we considered the forced vibration of a harmonic oscillator – that is, a system governed by the differential equation $mx'' + cx' + kx = f(t)$, for the case of the harmonic excitation $f(t) = F_0 \cos \Omega t$. Thus, besides a homogeneous solution we needed to find a particular solution, and that was done by the method of undetermined coefficients. The particular solution is especially important physically since even an infinitesimal amount of damping will cause the homogeneous solution to tend to zero as $t \to \infty$, so that the particular solution becomes the steady-state response. To understand the physical significance of that response we attached importance to the amplitude- and phase-response curves and discussed the phenomena of resonance and beats. Our discussion in this section has been limited in that we have considered only the case of harmonic excitation, whereas in applications $f(t)$ surely need not be harmonic. However, that case is important enough to deserve this special section. When we study the Laplace transform method in Chapter 5, we will be able to return to problems such as

$mx'' + cx' + kx = f(t)$ and, using the convenient Laplace transform methodology, obtain solutions for virtually any forcing function $f(t)$.

EXERCISES 3.8

1. Applying the initial conditions $x(0) = 0$ and $x'(0) = 0$ to (8), derive (13). Show that the same result can be obtained if we start with the form (9) instead of (8).

2. Derive (12) from (11).

3. (a) Derive (18). (b) Derive (19a,b,c).

4. The amplitude- and phase-response curves shown in Fig. 3 correspond to the equation $mx'' + kx = F_0 \cos \Omega t$. Obtain the equations of the analogous response curves for the equation $mx'' + kx = F_0 \sin \Omega t$, and give labeled sketches of the two curves.

5. Figure 6 shows the amplitude-response curves (E versus Ω) corresponding to (19b), for various values of c.

(a) What happens to the graph as $c \to \infty$? Is $E(\Omega)$ continuous on $0 \leq \Omega < \infty$ for $c = \infty$? Explain.
(b) From (19c), obtain the phase-response curves (Φ versus Ω), either by a careful freehand sketch or using a computer, for various values of c, being sure to include the important case $c = 0$. What happens to the graph as $c \to \infty$?

6. In Fig. 7 we show the approach of a representative response curve (solid) to the steady-state oscillation (dotted), for an underdamped system.

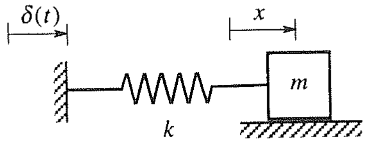
(a) Do the same (with a computer plot) for a critically damped case. The values of $m, c, k, F_0, \Omega, x(0), x'(0)$ are up to you, but the idea is to demonstrate graphically the approach to $x_p(t)$ clearly, as we have in Fig. 7.
(b) Same as (a), for an overdamped system, where $c = 4c_{cr}$, say.

7. Show that taking the limit of the response (13) as $\Omega \to \omega$, with the help of l'Hôpital's rule, does give (12), as claimed two paragraphs below (14).

8. Observe from Fig. 6 that the amplitude E tends to zero as $\Omega \to \infty$. Explain (physically, mathematically, or both) why that result makes sense.

9. (a) What choice of initial conditions $x(0)$ and $x'(0)$ will reduce the solution (20) to just the particular solution, $x(t) = E \cos(\Omega t + \Phi)$?
(b) Using a sketch of a representative $x_p(t)$ such as the dotted curve in Fig. 7, show the graphical significance of those special values of $x(0)$ and $x'(0)$.

10. Imagine the experimental means that would be required to apply a force $F_0 \cos \Omega t$ to a mass. It doesn't sound so hard if the mass is stationary, but imagine trying to apply such a force to a moving mass! In many physical applications, such as earthquake-induced vibration, the driving force is applied indirectly, by "shaking" the wall, rather than being applied directly to the mass. Specifically, for the system shown in the figure, use Newton's second law to show that if the wall is displaced laterally according to $\delta(t) = \delta_0 \cos \Omega t$, then the equation of motion of the mass m is $mx'' + kx = F_0 \cos \Omega t$, where $F_0 = k\delta_0$. Here, x and δ are measured relative to fixed points in space. NOTE: Observe that such an experiment is more readily performed since it is easier to apply a harmonic displacement $\delta(t)$ than a harmonic force; for instance, one could use a slider-crank mechanism (which converts circular motion to harmonic linear motion). Note further that a displacement input is precisely what an automobile suspension is subjected to when we drive over a bumpy road.



11. For the mechanical oscillator governed by the differential equation $mx'' + cx' + kx = F(t)$, obtain computer plots of the amplitude- and phase-response curves (E versus Ω and Φ versus Ω), for the case where $F(t) = 25 \sin \Omega t$, for these six values of the damping coefficient c: $0, 0.25c_{cr}, 0.5c_{cr}, c_{cr}, 2c_{cr}, 4c_{cr}$, where

(a) $m = 1, k = 1$
(b) $m = 2, k = 5$
(c) $m = 2, k = 10$
(d) $m = 4, k = 2$
(e) $m = 4, k = 10$

12. (*Complex function method*) Let L be a linear constant-coefficient differential operator, and consider the equation

$$L[x] = F_0 \cos \Omega t, \tag{12.1}$$

According to the method of undetermined coefficients, we can find a particular solution $x_p(t)$ by seeking $x_p(t) = A\cos\Omega t + B\sin\Omega t$ (or, in exceptional cases, t to an integer power times that). A slightly simpler line of approach that is sometimes used is as follows. Consider, in place of (12.1),

$$L[w] = F_0 e^{i\Omega t}, \tag{12.2}$$

Equation (12.2) is simpler than (12.1) in that to find a particular solution we need only one term, $w_p(t) = Ae^{i\Omega t}$. (If $z = a + ib$ is any complex number, it is standard to call $\operatorname{Re} z = a$ and $\operatorname{Im} z = b$ the **real part** and the **imaginary part** of z, respectively.) Because, according to Euler's formula, $e^{i\Omega t} = \cos\Omega t + i\sin\Omega t$, it follows that $\operatorname{Re} e^{i\Omega t} = \cos\Omega t$ and $\operatorname{Im} e^{i\Omega t} = \sin\Omega t$. Since the forcing function in (12.1) is the real part of the forcing function in (12.2), it seems plausible that $x_p(t)$ should be the real part of $w_p(t)$. Thus, we have the following method: to find a particular solution to (12.1) consider instead the simpler equation (12.2). Solve for $w_p(t)$ by seeking $w_p(t) = Ae^{i\Omega t}$, and then recover the desired $x_p(t)$ from $x_p(t) = \operatorname{Re} w_p(t)$.

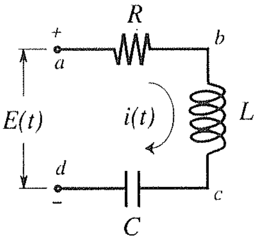
(a) Prove that the method described above works. HINT: The key is the linearity of L, so that if $w = u + iv$, then $L[w] = L[u + iv] = L[u] + iL[v]$.

(b)–(k) Use the method to obtain a particular solution to the given equation:

(b) $mx'' + cx' + kx = F_0 \cos\Omega t$
(c) $mx'' + cx' + kx = F_0 \sin\Omega t$
(d) $x' + 3x = 5\cos 2t$
(e) $x' - x = 4\sin 3t$
(f) $x'' - x' + x = \cos 2t$
(g) $x'' + 5x' + x = 3\sin 4t$
(h) $x'' - 2x' + x = 6\cos 5t$
(i) $x''' + x'' + x' + x = 3\sin t$
(j) $x'''' + x' + x = 3\cos t$
(k) $x'''' + 2x''' + 4x = 9\sin 6t$

13. (*Electrical circuit*) Recall from Section 2.3 that the equations governing the current $i(t)$ in the circuit shown, and the charge $Q(t)$ on the capacitor are

$$L\frac{d^2i}{dt^2} + R\frac{di}{dt} + \frac{1}{C}i = \frac{dE(t)}{dt}, \tag{13.1}$$



and

$$L\frac{d^2Q}{dt^2} + R\frac{dQ}{dt} + \frac{1}{C}Q = E(t), \tag{13.2}$$

respectively, where L, R, C, E, i, and Q are measured in henrys, ohms, farads, volts, amperes, and coulombs, respectively.

(a) Let $L = 2$, $R = 4$, and $C = 0.05$. Solve for $Q(t)$ subject to the initial conditions $Q(0) = Q'(0) = 0$, where $E(t) = 100$. Identify the steady-state solution. Give a computer plot of the solution for $Q(t)$ over a sufficiently long time period to clearly show the approach of Q to its steady state. (Naturally, all plots should be suitably labeled.)
(b) Same as (a), but for $C = 0.08$.
(c) Same as (a), but for $C = 0.2$.
(d) Same as (a), but for $E(t) = 10e^{-t}$.
(e) Same as (a), but for $E(t) = 10\,(1 - e^{-t})$.
(f) Same as (a), but for $E(t) = 50\left(1 + e^{-0.5t}\right)$.

3.9 Systems of Linear Differential Equations

Thus far we have considered the case where there is a single dependent variable such as the current $i(t)$ in a circuit, where t is the time. However, many problems involve two or more dependent variables. In the combustion of fossil fuels, for instance, there are many interacting chemical species (e.g., OH, CH_4, H, CO, H_2O, and so on) whose generation and demise, as a function of time, are governed by a large set of differential equations. A realistic model could easily contain 100 differential equations on 100 unknowns.

If there are two or more unknowns, then we are involved not with a single differential equation but with a system of such equations. For instance, according to the well known Lotka–Volterra model of predator-prey population dynamics, the populations $x(t)$ and $y(t)$ of predator and prey are governed by the system of two equations

$$\begin{aligned} x' &= -\alpha x + \beta xy, \\ y' &= \gamma y - \delta xy, \end{aligned} \tag{1}$$

where $\alpha, \beta, \gamma, \delta$ are empirical constants and t is the time. This particular system happens to be nonlinear because of the xy products; we will return to it in Chapter 7 when we study nonlinear systems. The present chapter is devoted exclusively to linear differential equations.

By definition, a **linear** first-order system of n equations in the n unknowns $x_1(t), \ldots, x_n(t)$ is of the form

$$\begin{aligned} a_{11}(t)x_1' + \cdots + a_{1n}(t)x_n' + b_{11}(t)x_1 + \cdots + b_{1n}(t)x_n &= f_1(t) \\ &\ \vdots \\ a_{n1}(t)x_1' + \cdots + a_{nn}(t)x_n' + b_{n1}(t)x_1 + \cdots + b_{nn}(t)x_n &= f_n(t), \end{aligned} \tag{2}$$

where the forcing functions $f_j(t)$ and the coefficients $a_{jk}(t)$ and $b_{jk}(t)$ are prescribed, and where it is convenient to use a double-subscript notation: $a_{jk}(t)$ denotes the coefficient of $x_k'(t)$ in the jth equation, and $b_{jk}(t)$ denotes the coefficient of $x_k(t)$ in the jth equation. We call (2) a first-order system because the highest derivatives are of first order. If the highest derivatives were of second order, we would call it a second-order system, and so on. A linear second-order system of n equations in the n unknowns $x_1(t), \ldots, x_n(t)$ would be of the same form as (2), but with each left-hand side being a linear combination of the second-, first-, and zeroth-order derivatives of the unkowns.

The system (2) is a generalization of the linear first-order equation $y' + p(x)y = q(x)$ in the one unknown $y(x)$ studied in Chapter 2. There, and in most of Chapters 1–3, we favored x as the generic independent variable and y as the generic dependent variable, but in this section the independent variable in most of our applications happens to be the time t, so we will use t as the independent variable.

As in the case of a single differential equation, by a **solution** of a system of differential equations (be they linear or not), in the unknowns $x_1(t), \ldots, x_n(t)$ over some t interval I, we mean a set of functions $x_1(t), \ldots, x_n(t)$ that reduce those equations to identities over I.

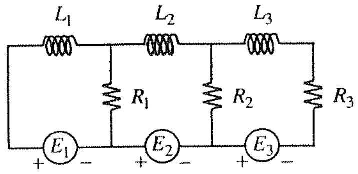


Figure 1. Circuit of Example 1.

3.9.1. Examples. let us begin by giving a few examples of how such systems arise in applications.

EXAMPLE 1. *RL Circuit.* Consider the circuit shown in Fig. 1, comprised of three

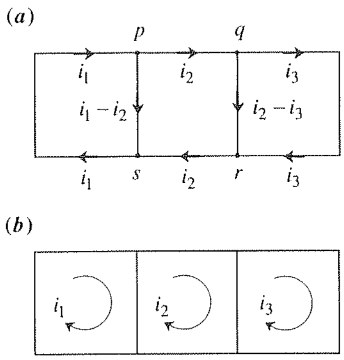


Figure 2. Current designations.

loops. We wish to obtain the differential equations governing the various currents in the circuit. There are two ways to proceed that are different but equivalent, and which correspond to the current labeling shown in Fig. 2a and 2b (in which we have omitted the circuit elements, for simplicity). First consider the former. If the current approaching the junction p from the "west" is designated as i_1 and the current leaving to the east is i_2, then it follows from Kirchoff's current law (namely, that the algebraic sum of the currents approaching or leaving any point of a circuit is zero) that the current to the south must be $i_1 - i_2$. Similarly, if we designate the current leaving the junction q to the east as i_3, then the current leaving to the south must be $i_2 - i_3$. With the current approaching r from the north and east being $i_2 - i_3$ and i_3, it follows that the current leaving to the west must be i_2. Similarly, the current leaving s to the west must be i_1.

Next, apply Kirchoff's voltage law (namely, that the algebraic sum of the voltage drops around each loop of the circuit must be zero) to each loop, recalling from Section 2.3 that the voltage drops across inductors, resistors, and capacitors (of which there are none in this particular circuit) are $L\dfrac{di}{dt}$, Ri, and $\dfrac{1}{C}\displaystyle\int i\,dt$, respectively. For the left-hand loop that step gives $L_1\dfrac{di_1}{dt} + R_1(i_1 - i_2) - E_1(t) = 0$, where the last term (corresponding to the applied voltage E_1) is counted as negative because it amounts to a voltage rise (according to the polarity denoted by the ± signs in Fig. 1) rather than a drop. Thus, we have for the left, middle, and right loops,

$$\begin{aligned}
&L_1 i_1' + R_1\,(i_1 - i_2) = E_1(t),\\
&L_2 i_2' + R_2\,(i_2 - i_3) + R_1\,(i_2 - i_1) = E_2(t),\\
&L_3 i_3' + R_3 i_3 + R_2\,(i_3 - i_2) = E_3(t),
\end{aligned} \tag{3}$$

respectively, or,

$$\begin{aligned}
&L_1 i_1' + R_1 i_1 - R_1 i_2 = E_1(t),\\
&L_2 i_2' - R_1 i_1 + (R_1 + R_2)\,i_2 - R_2 i_3 = E_2(t),\\
&L_3 i_3' - R_2 i_2 + (R_2 + R_3)\,i_3 = E_3(t),
\end{aligned} \tag{4}$$

where $E_1(t), E_2(t), E_3(t)$ are prescribed. It must be remembered that the currents do not need to flow in the directions assumed by the arrows; after all, they are the unknowns. If any of them turn out to be negative (at any given instant t), that merely means that they are flowing in the direction opposite to that tentatively assumed in Fig. 2a.

Alternatively, one can use the idea of "loop currents," as denoted in Fig. 2b. In that case the south-flowing currents in R_1 and R_2 are the net currents $i_1 - i_2$ and $i_2 - i_3$, respectively, just as in Fig. 2a. Either way, the result is the linear first-order system (4). ▮

It is important to see that the system (4) is *coupled*. That is, the individual equations contain more than one unknown so that we cannot separate them and solve the first for i_1 (for instance), the second for i_2, and the third for i_3. Put differently, the currents i_1, i_2, i_3 are interrelated. It is only natural for systems of differential equations to be coupled since the coupling is the mathematical expression of the relatedness of the dependent variables. On the other hand, if we write differential equations governing the current $i(t)$ in a circuit and the price of tea in China, $p(t)$, we would hardly expect those equations to be coupled and, indeed, it would hardly make sense to group them as part of the same system.

EXAMPLE 2. *LC Circuit.* For the circuit shown in Fig. 3, the same reasoning as above gives the integro-differential equations

$$\begin{aligned} \frac{1}{C_1}\int i_1\,dt + L\frac{d}{dt}(i_1 - i_2) &= E(t), \\ \frac{1}{C_2}\int i_2\,dt + L\frac{d}{dt}(i_2 - i_1) &= 0 \end{aligned} \tag{5}$$

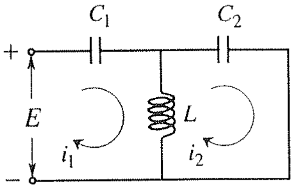


Figure 3. *LC* circuit.

on the currents $i_1(t)$ and $i_2(t)$ or, differentiating to eliminate the integral signs,

$$\begin{aligned} Li_1'' - Li_2'' + \frac{1}{C_1}i_1 &= E'(t), \\ Li_2'' - Li_1'' + \frac{1}{C_2}i_2 &= 0. \end{aligned} \tag{6}$$

Whereas (4) was a first-order system, (6) is of second order. ∎

EXAMPLE 3. *Mass-Spring System.* This time consider a mechanical system, shown in Fig. 4 and comprised of masses and springs. The masses rest on a frictionless table and are subjected to applied forces $F_1(t), F_2(t)$, respectively. When the displacements x_1 and x_2 are zero, the springs are neither stretched nor compressed, and we seek the equations of motion of the system, that is, the differential equations governing $x_1(t)$ and $x_2(t)$.

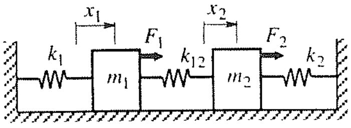


Figure 4. Mass-spring system.

The relevant physics is Newton's second law of motion, and Hooke's law for each of the three springs, as were discussed in Section 1.3. To proceed, it is useful to make a concrete assumption on x_1 and x_2. Specifically, suppose that at the instant t we have $x_1 > x_2 > 0$, as assumed in Fig. 5 (which figure, in the study of mechanics, is called a *free-body diagram*). Then the left spring is stretched by x_1 so it exerts a force to the left, on m_1, equal (according to Hooke's law) to k_1x_1. The middle spring is compressed by $x_1 - x_2$ so it exerts a force $k_{12}(x_1 - x_2)$ to the left on m_1 and to the right on m_2, and the right spring is compressed by x_2 and exerts a force k_2x_2 to the left on m_2, as shown in the figure. With the help of the information given in Fig. 5, Newton's second law for each of the two masses gives

$$\begin{aligned} m_1x_1'' &= -k_1x_1 - k_{12}(x_1 - x_2) + F_1(t), \\ m_2x_2'' &= -k_2x_2 + k_{12}(x_1 - x_2) + F_2(t) \end{aligned} \tag{7}$$

as the desired equations of motion – or, rearranging terms,

$$\begin{aligned} m_1x_1'' + (k_1 + k_{12})x_1 - k_{12}x_2 &= F_1(t), \\ m_2x_2'' - k_{12}x_1 + (k_2 + k_{12})x_2 &= F_2(t). \end{aligned} \tag{8}$$

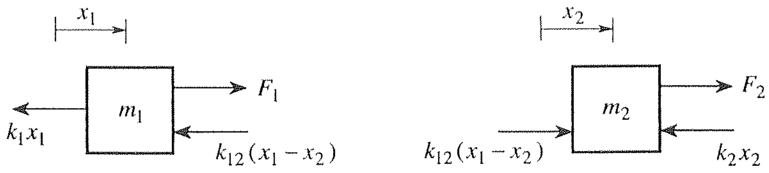


Figure 5. Free-body diagram of the masses.

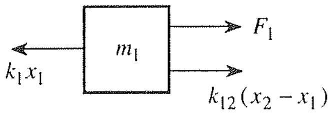


Figure 6. Revised free-body diagram for m_1.

COMMENT. Our assumption that $x_1 > x_2 > 0$ was only for definiteness; the resulting equations (8) are insensitive to whatever such assumption is made. For instance, suppose that we assume, instead, that $x_2 > x_1 > 0$. Then the middle spring is stretched by $x_2 - x_1$, so the free-body diagram of m_1 changes to that shown in Fig. 6, and Newton's law for m_1 gives $m_1x_1'' = -k_1x_1 + k_{12}(x_2 - x_1) + F_1(t)$, which is seen to be equivalent to the first of equations (7). Similarly for m_2. ▮

3.9.2. Existence and uniqueness. The fundamental theorem regarding existence and uniqueness is as follows.*

THEOREM 3.9.1 *Existence and Uniqueness for Linear First-Order Systems*
Let the functions $a_{11}(t), a_{12}(t), \dots, a_{nn}(t)$ and $f_1(t), \dots, f_n(t)$ be continuous on a closed interval I. And let numbers $b_1, \dots, b_n$ be given such that

$$x_1(a) = b_1, \quad x_2(a) = b_2, \quad \cdots, \quad x_n(a) = b_n, \tag{9}$$

where a is a given point in I. Then the system

$$\begin{aligned} x_1' &= a_{11}(t)x_1 + a_{12}(t)x_2 + \cdots + a_{1n}(t)x_n + f_1(t), \\ &\vdots \\ x_n' &= a_{n1}(t)x_1 + a_{n2}(t)x_2 + \cdots + a_{nn}(t)x_n + f_n(t), \end{aligned} \tag{10}$$

*There is a subtle point that is worth noting, namely, that (10) is not quite of the same form as the general first-order linear system (2) in that its left-hand sides are simply $x_1', \dots, x_n'$ rather than linear combinations of those terms. (What follows presumes that you have already studied the sections on matrices, rank, and Gauss–Jordan reduction.) The idea is that (2) can be reduced to the form (10) by elementary row operations, as in the Gauss–Jordan reduction of linear algebraic equations – unless the rank of the $\{a_{jk}(t)\}$ matrix is less than n. In that case, such operations would yield at least one equation, at the bottom of the system, which has no derivatives in it. If not all of the coefficients of the undifferentiated x_j terms in that equation are zero, then one could use that equation to solve for one of the x_j's in terms of the others and use that result to reduce the system by one unknown and one equation; if all of the coefficients of the undifferentiated x_j terms in that equation are zero, then that equation would either be $0 = 0$, which could be discarded, or zero equal to some nonzero prescribed function of t, which would cause the system to have no solution. To avoid these singular cases, it is conventional to use the form (10), rather than (2), in the existence and uniqueness theorem.

subject to the initial conditions (9), has a unique solution on the entire interval I.

Observe that we have added the word "entire" for emphasis, for recall from Section 2.4 that the Existence and Uniqueness Theorem 2.4.1 for the *non*linear initial-value problem $y'(x) = f(x, y)$ with initial condition $y(a) = b$ is a local one; it guarantees the existence of a unique solution over *some* interval $|x - a| < h$, but it does not tell us how big h can be. In contrast, Theorem 3.9.1 tells us that the solution exists and is unique over the entire interval I over which the specified conditions are met.

EXAMPLE 4. *Example 1, Revisited.* Suppose that we add initial conditions, say $i_1(0) = b_1, i_2(0) = b_2, i_3(0) = b_3$ to the system (4) governing the RL circuit of Example 1. If the $E_j(t)$'s are continuous on $0 \le t \le T$ and the L_j's are nonzero [so we can divide through by them in reducing (4) to the form of (10)] then, according to Theorem 3.9.1, the initial-value problem has a solution on $0 \le t \le T$, and it is unique. ∎

It would appear that Theorem 3.9.1 does not apply to the system (8) of Example 3 because the latter is of second order rather than first. However, and this is important, higher-order systems can be reduced to first-order ones by introducing artificial, or auxiliary, dependent variables.

EXAMPLE 5. Reduce the second-order system (8) to a first-order system. The idea is to introduce artificial dependent variables u and v according to $x_1' = u$ and $x_2' = v$ because then the second-order derivatives x_1'' and x_2'' become first-order derivatives u' and v', respectively. Thus, (8) can be re-expressed, equivalently, as the first-order system

$$\begin{aligned} x_1'(t) &= u, \\ u'(t) &= -\frac{k_1 + k_{12}}{m_1} x_1 + \frac{k_{12}}{m_1} x_2 + \frac{1}{m_1} F_1(t), \\ x_2'(t) &= v, \\ v'(t) &= \frac{k_{12}}{m_2} x_1 - \frac{k_2 + k_{12}}{m_2} x_2 + \frac{1}{m_2} F_2(t). \end{aligned} \tag{11}$$

To see that this system is of the form (10), let "x_1" $= x_1$, "x_2" $= u$, "x_3" $= x_2$, and "x_4" $= v$. Then $a_{11} = a_{13} = a_{14} = 0$, $a_{12} = 1$, $f_1(t) = 0$, $a_{21} = -(k_1 + k_{12})/m_1$, $a_{22} = a_{44} = 0$, $a_{23} = k_{12}/m_1$, $f_2(t) = F_1(t)/m_1$, and so on. All of the $a_{jk}(t)$ coefficients are constants and hence continuous for all t. Let the forcing functions $F_1(t)$ and $F_2(t)$ be continuous on $0 \le t < \infty$.

Thus, according to Theorem 3.9.1, if we prescribe initial conditions $x_1(0)$, $u(0)$, $x_2(0)$, $v(0)$, then the initial-value problem consisting of (11), together with those initial conditions, will have a unique solution for $x_1(t), u(t), x_2(t), v(t)$. Equivalently, the initial-value problem consisting of (8), together with prescribed initial values $x_1(0)$, $x_1'(0)$, $x_2(0)$, $x_2'(0)$, will have a unique solution for $x_1(t), x_2(t)$. ∎

Consider one more example on auxiliary variables.

EXAMPLE 6. Consider the third-order equation

$$x''' + 2tx'' - x' + (\sin t)x = \cos t, \tag{12}$$

which *is* a system, a system of n equations in n unknowns, where $n = 1$. To reduce it to a system of first-order equations, introduce auxiliary variables u, v according to $x' = u$ and $x'' = u' = v$. Then

$$\begin{aligned} x' &= u, \\ u' &= v, \\ v' &= -(\sin t)x + u - 2tv + \cos t \end{aligned} \tag{13}$$

is the desired equivalent first-order system, where the last of the three equations follows from the first two together with (12). ▮

3.9.3. Solution by elimination. We now give a method of solution of systems of linear differential equations for the special case of constant coefficients, a method of elimination that is well suited to systems that are small enough for us to carry out the steps by hand.

We introduce the method with an example after first recalling (from Section 3.3) the idea of a linear differential operator,

$$\begin{aligned} L &= a_0(t)\frac{d^n}{dt^n} + a_1(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_n(t) \\ &= a_0(t)D^n + a_1(t)D^{n-1} + \cdots + a_n(t), \end{aligned}$$

where D denotes $\dfrac{d}{dt}$, D^2 denotes $\dfrac{d^2}{dt^2}$, and so on. By $L[x]$ we mean the function $a_0\dfrac{d^n x}{dt^n} + a_1\dfrac{d^{n-1}x}{dt^{n-1}} + \cdots + a_n x$. We say that L is of order n (if a_0 is not identically zero) and that it "acts" on x, or "operates" on x. Further, by $L_1L_2[x]$ we mean $L_1[L_2[x]]$; that is, first the operator immediately to the left of x acts on x, then the operator to the left of that acts on the result. Two operators, say L_1 and L_2, are said to be equal if $L_1[x] = L_2[x]$ for all functions $x(t)$ (that are sufficiently differentiable for L_1 and L_2 to act on them). Finally, *in general, differential operators do not commute*: $L_1L_2 \neq L_2L_1$. For instance, if $L_1 = D$ and $L_2 = tD$, then $L_1L_2[x] = D(tDx) = D(tx') = tx'' + x'$, whereas $L_2L_1[x] = tD(Dx) = tDx' = tx''$. However, they *do commute if their a_j coefficients are constants*. For instance,

$$(2D - 1)(D + 3)x = (2D - 1)(x' + 3x) = 2x'' + 6x' - x' - 3x,$$

and

$$(D + 3)(2D - 1)x = (D + 3)(2x' - x) = 2x'' - x' + 6x' - 3x$$

are identical for all functions $x(t)$, so $(2D-1)(D+3)=(D+3)(2D-1)$.

EXAMPLE 7. To solve the system

$$x'-x-y=3t, \tag{14a}$$
$$x'+y'-5x-2y=5, \tag{14b}$$

it is convenient to begin by re-expressing it as

$$(D-1)x-y=3t, \tag{15a}$$
$$(D-5)x+(D-2)y=5, \tag{15b}$$

or

$$L_1[x]+L_2[y]=3t, \tag{16a}$$
$$L_3[x]+L_4[y]=5, \tag{16b}$$

where $L_1=D-1, L_2=-1$, and so on. To solve by the method of elimination, let us operate on (16a) with L_3 and on (16b) with L_1, giving

$$L_3L_1[x]+L_3L_2[y]=L_3[3t], \tag{17a}$$
$$L_1L_3[x]+L_1L_4[y]=L_1[5], \tag{17b}$$

where we have used the linearity of L_3 in writing $L_3\,[L_1[x]+L_2[y]]$ as $L_3L_1[x]+L_3L_2[y]$ in obtaining (17a) and, similarly, the linearity of L_1 in obtaining (17b). Subtracting one equation from the other, and cancelling the x terms because $L_3L_1=L_1L_3$, enables us to eliminate x and to obtain the equation

$$(L_1L_4-L_3L_2)\,[y]=L_1[5]-L_3[3t] \tag{18}$$

on y alone. At this point we can return to the non-operator form, with $L_1L_4-L_3L_2=(D-1)(D-2)-(D-5)(-1)=D^2-2D-3$ and $L_1[5]-L_3[3t]=(D-1)(5)-(D-5)(3t)=15t-8$. Thus,

$$y''-2y'-3y=15t-8, \tag{19}$$

which admits the general solution

$$y(t)=Ae^{3t}+Be^{-t}-5t+6. \tag{20}$$

To find $x(t)$, we can proceed in the same manner. This time, operate on (16a) with L_4 and on (16b) with L_2:

$$L_4L_1[x]+L_4L_2[y]=L_4[3t] \tag{21a}$$
$$L_2L_3[x]+L_2L_4[y]=L_2[5], \tag{21b}$$

and subtraction gives

$$(L_4L_1-L_2L_3)\,[x]=L_4[3t]-L_2[5], \tag{22}$$

or

$$x'' - 2x' - 3x = 8 - 6t, \tag{23}$$

with general solution

$$x(t) = Ce^{3t} + Ee^{-t} + 2t - 4. \tag{24}$$

(We avoid using D as an integration constant because $D = d/dt$ here.)

It might appear that A, B, C, E are all arbitrary, but don't forget that x and y are related through (14), so these constants might be related as well. In fact, putting (20) and (24) into (14a) gives, after cancellation of terms,

$$(2C - A)e^{3t} - (2E + B)e^{-t} = 0, \tag{25}$$

and the linear independence of e^{3t} and e^{-t} requires that $A = 2C$ and $B = -2E$. Putting (20) and (24) into (14b) gives this same result.

Thus, the general solution of (14) is

$$x(t) = Ce^{3t} + Ee^{-t} + 2t - 4, \tag{26a}$$
$$y(t) = 2Ce^{3t} - 2Ee^{-t} - 5t + 6. \tag{26b}$$

COMMENT 1. With hindsight, it would have been easier to eliminate y first and solve for x since we could have put that x [namely, as given by (26a)] into (14a) and solved that equation for y. That step would have produced (26b) directly.

COMMENT 2. Notice that (14) is not of the "standard" form (10) because (14b) has both x' and y' in it. While we need it to be in that form to apply Theorem 3.9.1, we do not need the system to be of that form to apply the method of elimination. ▮

A review of the steps in the elimination process reveals that the operators $L_1, \dots, L_4$ might just as well have been constants, by the way we have manipulated them. In fact, a useful way to organize the procedure is to use Cramer's rule (Section 10.6). For instance, if we have two differential equations

$$L_1[x] + L_2[y] = f_1(t), \tag{27a}$$
$$L_3[x] + L_4[y] = f_2(t), \tag{27b}$$

we can, heuristically, use Cramer's rule to write

$$x = \frac{\begin{vmatrix} f_1 & L_2 \\ f_2 & L_4 \end{vmatrix}}{\begin{vmatrix} L_1 & L_2 \\ L_3 & L_4 \end{vmatrix}} = \frac{L_4[f_1] - L_2[f_2]}{L_1 L_4 - L_2 L_3}, \tag{28a}$$

$$y = \frac{\begin{vmatrix} L_1 & f_1 \\ L_3 & f_2 \end{vmatrix}}{\begin{vmatrix} L_1 & L_2 \\ L_3 & L_4 \end{vmatrix}} = \frac{L_1[f_2] - L_3[f_1]}{L_1 L_4 - L_2 L_3}. \tag{28b}$$

Of course, the division by an operator on the right-hand sides of (28a,b) is not defined, so we need to put the $L_1L_4 - L_2L_3$ back up on the left-hand side, where it came from. That step gives

$$(L_1L_4 - L_2L_3)[x] = L_4[f_1] - L_2[f_2] \tag{29a}$$

$$(L_1L_4 - L_2L_3)[y] = L_1[f_2] - L_3[f_1], \tag{29b}$$

which equations correspond to (22) and (18), respectively, in Example 7. Again, this approach is heuristic, but it does give the correct result and is readily applied – and extended to systems of three equations in three unknowns, four in four unknowns, and so on.

What might possibly go wrong with our foregoing solution of (27)? In the application of Cramer's rule to linear algebraic equations, the case where the determinant in the denominator vanishes is singular, and either there are no solutions (the system is "inconsistent") or there is an infinite number of them (the system is "redundant"). Likewise, the system (27) is singular if $L_1L_4 - L_2L_3$ is zero and is either inconsistent (with no solution) or redundant (with infinitely many linearly independent solutions). For instance, the system

$$Dx + 2Dy = 1, \tag{30a}$$

$$2Dx + 4Dy = 3 \tag{30b}$$

has $L_1L_4 - L_2L_3 = 4D^2 - 4D^2 = 0$ and has no solution since the left-hand sides are in the ratio 1:2, whereas the right-hand sides are in the ratio 1:3. However, if we change the 3 to a 2, then the new system still has $L_1L_4 - L_2L_3 = 0$ but is now consistent. Indeed, then the second equation is merely twice the first and can be discarded, leaving the single equation $Dx + 2Dy = 1$ in the two unknowns $x(t)$ and $y(t)$. We can choose one of these arbitrarily and use $Dx + 2Dy = 1$ to solve for the other, so there are infinitely many linearly independent solutions. Understand that the singular nature of (30), and the modified system, is intrinsic to those systems and is not a fault of the method of elimination.

In the generic case, however, $L_1L_4 - L_2L_3 \neq 0$ and we can solve (29a) and (29b) for $x(t)$ and $y(t)$, respectively. It can be shown* that the number of independent arbitrary integration constants is the same as the degree of the determinantal polynomial $L_1L_4 - L_2L_3$. In Example 7, for instance, $L_1L_4 - L_2L_3 = D^2 - 2D - 3$ is of second degree, so we could have known in advance that there would be two independent arbitrary constants.

EXAMPLE 8. *Mass-Spring System in Fig. 4.* Let us study the two-mass system shown in Fig. 4, and let $m_1 = m_2 = k_1 = k_{12} = k_2 = 1$ and $F_1(t) = F_2(t) = 0$, for definiteness. Then equations (8) become

*See pages 144–150 in the classic treatise by E. L. Ince, *Ordinary Differential Equations* (New York: Dover, 1956).

$$(D^2+2)\,x_1 - x_2 = 0, \tag{31a}$$
$$-x_1 + (D^2+2)\,x_2 = 0. \tag{31b}$$

With $L_1 = L_4 = D^2 + 2$ and $L_2 = L_3 = -1$, and $f_1(t) = f_2(t) = 0$, (29a,b) become

$$(D^4 + 4D^2 + 3)\,x_1 = 0, \tag{32a}$$
$$(D^4 + 4D^2 + 3)\,x_2 = 0, \tag{32b}$$

so (Exercise 2)

$$x_1(t) = A\cos t + B\sin t + C\cos\sqrt{3}\,t + E\sin\sqrt{3}\,t, \tag{33a}$$
$$x_2(t) = F\cos t + G\sin t + H\cos\sqrt{3}\,t + I\sin\sqrt{3}\,t. \tag{33b}$$

To determine any relationships among the constants $A, B, \dots, I$, we put (33) into (31a) [or (32b), the result would be the same] and find that

$$(A-F)\cos t + (B-G)\sin t - (C+H)\cos\sqrt{3}\,t - (E+I)\sin\sqrt{3}\,t = 0,$$

from which we learn that $F = A$, $G = B$, $H = -C$, and $I = -E$, so the general solution of (31) is

$$x_1(t) = A\cos t + B\sin t + C\cos\sqrt{3}\,t + E\sin\sqrt{3}\,t, \tag{34a}$$
$$x_2(t) = A\cos t + B\sin t - C\cos\sqrt{3}\,t - E\sin\sqrt{3}\,t. \tag{34b}$$

The determinantal polynomial was of fourth degree and, as asserted above, there are four independent arbitrary integration constants. There are important things to say about the result expressed in (34):

COMMENT 1. It will be more illuminating to re-express (34) in the form

$$x_1(t) = G\sin(t+\phi) + H\sin\left(\sqrt{3}\,t + \psi\right), \tag{35a}$$
$$x_2(t) = G\sin(t+\phi) - H\sin\left(\sqrt{3}\,t + \psi\right), \tag{35b}$$

where the four constants G, H, ϕ, ψ are determined from the initial conditions $x_1(0)$, $x_1'(0)$, $x_2(0)$, and $x_2'(0)$. While neither $x_1(t)$ nor $x_2(t)$ is a pure sinusoid, each is a superposition of two pure sinusoids, the frequencies of which are characteristics of the system (i.e., independent of the initial conditions). Those frequencies, $\omega = 1$ rad/sec and $\omega = \sqrt{3}$ rad/sec, are the **natural frequencies** of the system. If the initial conditions are such that $H = 0$, then the motion is of the form

$$x_1(t) = G\sin(t+\phi), \quad x_2(t) = G\sin(t+\phi); \tag{36}$$

that is, the two masses swing in unison at the lower frequency $\omega = 1$. Such a motion is called a **low mode** motion because it is at the lower of the two natural frequencies. If instead the initial conditions are such that $G = 0$, then

$$x_1(t) = H\sin(\sqrt{3}\,t + \psi), \quad x_2(t) = -H\sin(\sqrt{3}\,t + \psi); \tag{37}$$

the masses swing in opposition, at the higher frequency $\omega = \sqrt{3}$ rad/sec, so the latter is called a **high mode** motion. For instance, the initial conditions $x_1(0) = x_2(0) = 1$ and $x_1'(0) = x_2'(0) = 0$ give (Exercise 7) the purely low mode motion

$$\begin{aligned} x_1(t) &= \sin(t + \pi/2) = \cos t, \\ x_2(t) &= \sin(t + \pi/2) = \cos t, \end{aligned} \tag{38}$$

and the conditions $x_1(0) = 1$, $x_2(0) = -1$, and $x_1'(0) = x_2'(0) = 0$ give the purely high mode motion

$$\begin{aligned} x_1(t) &= \sin(\sqrt{3}\,t + \pi/2) = \cos\sqrt{3}\,t, \\ x_2(t) &= -\sin(\sqrt{3}\,t + \pi/2) = -\cos\sqrt{3}\,t. \end{aligned} \tag{39}$$

If, instead, $x_1(0) = 1$ and $x_2(0) = x_1'(0) = x_2'(0) = 0$, say, then both G and H will be nonzero and the motion will be a linear combination of the low and high modes.

COMMENT 2. Why is the frequency corresponding to the masses swinging in opposition higher than that corresponding to the masses swinging in unison? Remember from the single-mass case studied in Section 3.5 that the natural frequency in that case is $\sqrt{k/m}$; that is, the stiffer the system (the larger the value of k), the higher the frequency. For the two-mass system, observe that in the low mode the middle spring is completely inactive, whereas in the high mode it is being stretched and compressed. Thus, there is more stiffness encountered in the high mode, so the high mode frequency is higher.

COMMENT 3. Just as the free vibration of a single mass is governed by one differential equation, $mx'' + kx = 0$, and has a single mode of vibration with natural frequency $\omega = \sqrt{k/m}$, a two-mass system is governed by two differential equations and its general vibration is a linear combination of two modes (unison and opposition in this example), each with its own natural frequency. Similarly, the free vibration of an n-mass system will be governed by n differential equations, and its general vibration will be a linear combination of n distinct modes, each with its own pattern and natural frequency. In the limit, we can think of a continuous system, such as a beam, as an infinite-mass system, an infinite number of tiny masses connected together. In that limiting case, in place of an infinite number of ordinary differential equations we obtain a *partial* differential equation on the deflection $y(x, t)$, solution of which yields the general solution as a linear combination of an infinite number of discrete modes of vibration. In applications it is important to know the natural frequencies of a given system because if it is driven by a harmonic forcing function, then it will have a large, perhaps catastrophic, response if the driving frequency is close to one of the natural frequencies.

COMMENT 4. Finally, we note that molecules and atoms can be modeled as mass-spring systems, and the spectrum of the natural frequencies are of great importance in determining their allowable energy levels. ▮

We will have more to say about the foregoing example later, when we study matrix theory and the eigenvalue problem.

Observe that once a system of linear constant coefficient equations is converted by the process of elimination to a set of uncoupled equations such as (32a,b), the homogeneous solutions of those equations can be sought in the usual exponential form. In fact, one can do that even at the outset, without first going through the process of elimination. For instance, to solve (31a,b) one can start out by seeking a solution in the form $x_1(t) = \xi_1 e^{rt}$ and $x_2(t) = \xi_2 e^{rt}$. Putting those forms into (31a,b) gives what is known as an *eigenvalue problem* on the unknown constants ξ_1, ξ_2 and r. That discussion is best reserved for the chapters on matrix theory and linear algebra, as an important application of the eigenvalue problem, so we will not pursue it in the present section.

Closure. Systems of ordinary differential equations arise in the modeling of physical systems that involve more than one dependent variable. For instance, in modeling an ecological system such as the fish populations in a given lake, the dependent variables might be the populations of each fish species, as a function of the independent variable t. Surely these populations are interrelated (for instance, one species might be the primary food supply for another species), so the governing differential equations will be coupled. It is precisely the coupling that produces the interest in this section because if they are not coupled then we can solve them, individually, by the methods developed in preceding sections.

Our first step was to give the basic existence and uniqueness theorem. That theorem guaranteed both existence and uniqueness, under rather mild conditions of continuity, over an interval that is known in advance. The theorem applied to first-order systems, but we showed that systems of higher order can be converted to first-order systems by suitable introduction of auxiliary dependent variables.

Then we outlined a method of elimination for systems with constant coefficients. Elimination is similar to the steps in the solution of linear algebraic equations by Gauss elimination, where the coefficients of the unknowns are operators rather than numbers. The correct result can even be obtained by using Cramer's rule, provided that the determinantal operator in the denominator does not vanish, and provided that we move that operator back "upstairs" – as we did in converting (28) to (29). If the operator does vanish, then the problem is singular and there will be no solution or infinitely many linearly independent solutions.

In subsequent chapters on matrix theory we shall return to systems of linear differential equations with constant coefficients and develop additional solution techniques that are based upon the so-called eigenvalue problem.

Computer software. Often, one can solve systems of differential equations using computer-algebra systems. For instance, to find the general solution of the system

$$\begin{aligned}(D+1)x + 2y &= 0,\\ 3x + (D+2)y &= 0\end{aligned}$$

using *Maple*, enter

$$\begin{aligned}&\text{dsolve}(\{\text{diff}(x(t),t) + x(t) + 2 * y(t) = 0,\ 3 * x(t) + \text{diff}(y(t),t)\\ &+2 * y(t) = 0\}, \{x(t), y(t)\});\end{aligned}$$

and return. The result is the general solution

$$\{\, y(t) = \_C2 \exp(t) + 3/2\, \_C1 \exp(-4t), \\ x(t) = -\_C2 \exp(t) + \_C1 \exp(-4t)\, \}$$

If we wish to include initial conditions $x(0) = 3$, $y(0) = 2$, use instead the command

$$\text{dsolve}(\{\text{diff}(x(t),t) + x(t) + 2 * y(t) = 0,\ 3 * x(t) + \text{diff}(y(t),t) + 2 * y(t) = 0, \\ x(0) = 3,\ y(0) = 2\},\ \{x(t), y(t)\});$$

The result is the particular solution

$$\{y(t) = -\exp(t) + 3\exp(-4t),\ \ x(t) = \exp(t) + 2\exp(-4t)\}$$

Alternatively, one can first define the two equations and then call them in the dsolve command. That is, enter

$$\text{deq1} := \text{diff}(x(t),t) + x(t) + 2 * y(t) = 0 : \\ \text{deq2} := 3 * x(t) + \text{diff}(y(t),t) + 2 * y(t) = 0 :$$

The colon at the end of each line indicates that it is a definition, not a command. Commands are followed by semicolons. Now, enter the dsolve command:

$$\text{dsolve}(\{\text{deq1, deq2},\ x(0) = 3,\ y(0) = 2\},\ \{x(t), y(t)\});$$

and return. The result is

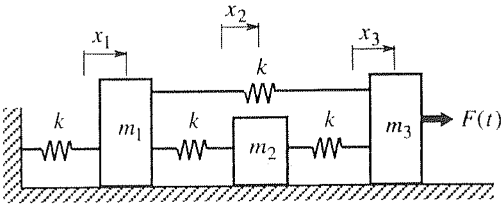
$$\{y(t) = -\exp(t) + 3\exp(-4t),\ \ x(t) = \exp(t) + 2\exp(-4t)\}$$

EXERCISES 3.9

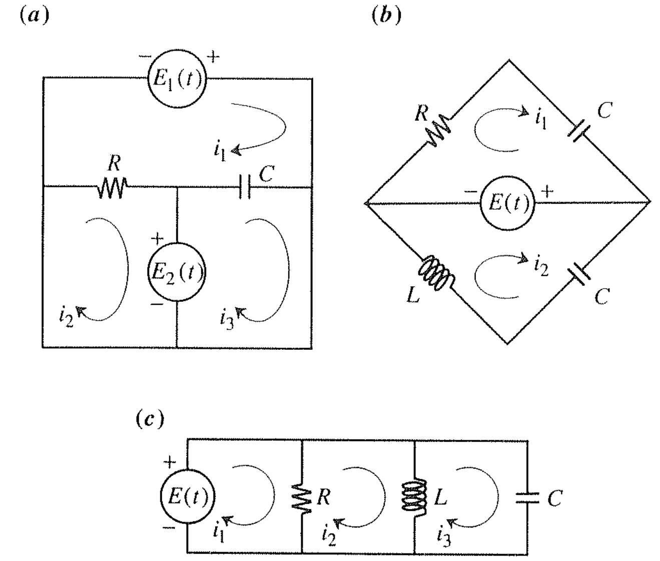
1. Derive the solution (20) of (19).

2. Derive the solutions (33a,b) of (32a,b).

3. Derive the system of differential equations governing the displacements $x_j(t)$, using the assumption that $x_1 > x_2 > x_3 > 0$. Repeat the derivation assuming instead that $x_3 > x_2 > x_1 > 0$ and again, assuming that $x_1 > x_3 > x_2 > 0$, and show that the resulting equations are the same, independent of these different assumptions.



4.(a),(b),(c) Derive the system of differential equations governing the currents $i_j(t)$, but you need not solve them. State any physical laws that you use.



5. Obtain the general solution by the method of elimination, either step-by-step or using the Cramer's rule shortcut.

(a) $(D-1)x + Dy = 0$
$(D+1)x + (2D+2)y = 0$

(b) $(D-1)x + 2Dy = 0$
$(D+1)x + 4Dy = 0$

(c) $Dx + (D-1)y = 5$
$2(D+1)x + (D+1)y = 0$

(d) $x' + y' = y + t$
$x' - 3y' = -x + 2$

(e) $x' = \sin t - y$
$y' = -9x + 4$

(f) $x' = x - 8y$
$y' = -x - y - 3t^2$

(g) $x' = 2x + 6y - t + 7$
$y' = 2x - 2y$

(h) $2x' + y' + x + y = t^2 - 1$
$x' + y' + x + y = 0$

(i) $x' + y' + x - y = e^t$
$x' + 2y' + 2x - 2y = 1 - t$

(j) $x' - 3x + y = 4\sin 2t$
$3x + y' - y = 6$

(k) $x'' = x - 4y$
$y'' = -2x - y$

(l) $x'' = x - 2y$
$y'' = 2x - 4y$

(m) $x'' - x + 2y = 0$
$-2x + y'' + 4y = 1 - t^2$

(n) $x'' - x + 3y = 0$
$y'' + x + y = 4$

(o) $x'' + x + y = 24$
$y'' + 3x - y = -8$

(p) $x'' + y'' = x$
$3x'' - y'' = y + 6$

(q) $(2D^2 + 3)x + (2D + 1)y = 4e^{3t} - 7$
$Dx + (D - 2)y = 2$

6. (a)–(q) Find the general solution of the corresponding problem in Exercise 5 using computer software. Separately, make up any set of initial conditions, and use the computer to find the particular solution corresponding to those initial conditions.

7. (*Mass-spring system of Examples 3 and 8*) (a) Derive the particular solutions (38) and (39) from the general solution (35) by applying the given sets of initial conditions.
(b) Evaluate G, H, ϕ, ψ for the initial conditions $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$, and show that both modes are present in the solution. Obtain a computer plot of $x_1(t)$ and $x_2(t)$, over $0 \le t \le 20$ (so as to show several cycles).
(c) Same as (b), for $x_1'(0) = 1, x_1(0) = x_2(0) = x_2'(0) = 0$.
(d) Same as (b), for $x_1(0) = x_2(0) = 0, x_1'(0) = 2, x_2'(0) = 3$.
(e) Same as (b), for $x_1(0) = x_2(0) = 0, x_1'(0) = 2$, $x_2'(0) = -1$.

8. (*Chemical kinetics*) Two substances, with concentrations $x(t)$ and $y(t)$, react to form a third substance, with concentration $z(t)$. The reaction is governed by the system $x' + \alpha x = 0, z' = \beta y$ and $x + y + z = \gamma$, where α, β, γ are known positive constants. Solve for $x(t)$, $y(t)$, $z(t)$, subject to the initial conditions $z(0) = z'(0) = 0$ for these cases:

(a) $\alpha \neq \beta$
(b) $\alpha = \beta$ HINT: Apply l'Hôpital's rule to your answer to part (a).

9. (*Motion of a charged mass*) Consider a particle of mass m, carrying an electrical charge q, and moving in a uniform magnetic field of strength B. The field is in the positive z direction. The equations of motion of the particle are

$$\begin{aligned} mx'' &= qBy', \\ my'' &= -qBx', \\ mz'' &= 0, \end{aligned} \tag{9.1}$$

where $x(t), y(t), z(t)$ are the x, y, z displacements as a function of the time t.

(a) Find the general solution of (9.1) for $x(t), y(t), z(t)$. How many independent arbitrary constants of integration are there?

(b) Show that by a suitable choice of initial conditions the motion can be a circle in the x, y plane, of any desired radius R and centered at any desired point x_0, y_0. Propose such a set of initial conditions.
(c) Besides a circular motion in a constant z plane, are any other types of motion possible? Explain.

10. Show that the given system is singular (i.e., either inconsistent or redundant). If it has no solutions show that; if it has solutions find them.

(a) $x' - x + y = t$
$x'' + y' - x + y = t^2$

(b) $(D - 1)x + y = t$
$(D^2 - 1)x + (D + 1)y = t + 1$

(c) $(D + 1)x - Dy = e^t$
$(D^2 - 1)x - (D^2 - D)y = 0$

(d) $(D + 1)x + Dy = e^t$
$(D^2 - 1)x + (D^2 - D)y = 3t$

Chapter 3 Review

A differential equation is far more tractable, insofar as analytical solution is concerned, if it is linear than if it is nonlinear. We could see a hint of that even in Chapter 2, where we were able to derive the general solution of the general first-order linear equation $y' + p(x)y = q(x)$ but had success with nonlinear equations only in special cases. In fact, for linear equations of any order (first, second, or higher) a number of important results follow.

The most important is that for an nth-order *linear* differential equation $L[y] = f(x)$, with constant coefficients or not, a general solution is expressible as the sum of a general solution $y_h(x)$ to the homogeneous equation $L[y] = 0$, and any particular solution $y_p(x)$ to the full equation $L[y] = f$:

$$y(x) = y_h(x) + y_p(x).$$

In turn, $y_h(x)$ is expressible as an arbitrary linear combination of any n LI (linearly independent) solutions of $L[y] = 0$:

$$y_h(x) = C_1 y_1(x) + \cdots + C_n y_n(x).$$

Thus, linear independence is introduced early, in Section 3.2, and theorems are provided for testing a given set of functions to see if they are LI or not. We then show how to find the solutions $y_1(x), \ldots, y_n(x)$ for the following two extremely important cases: for constant-coefficient equations and for Cauchy–Euler equations.

For *constant-coefficient equations* the idea is to seek $y_h(x)$ in the exponential form $e^{\lambda x}$. Putting that form into $L[y] = 0$ gives an nth-degree polynomial equation on λ, called the *characteristic equation*. Each nonrepeated root λ_j contributes a solution $e^{\lambda_j x}$, and each repeated root λ_j of order k contributes k solutions $e^{\lambda_j x}, xe^{\lambda_j x}, \ldots, x^{k-1}e^{\lambda_j x}$.

For *Cauchy–Euler equations* the form $e^{\lambda x}$ does *not* work. Rather, the idea is to seek $y_h(x)$ in the power form x^{λ}. Each nonrepeated root λ_j contributes a solution x^{λ_j}, and each repeated root λ_j of order k contributes k solutions $x^{\lambda_j}, (\ln x)x^{\lambda_j}, \ldots, (\ln x)^k x^{\lambda_j}$.

Two different methods are put forward for finding particular solutions, the method of *undetermined coefficients* and Lagrange's method of *variation of parameters*. Undetermined coefficients is easier to apply but is subject to the conditions that

(i) besides being linear, L must be of constant-coefficient type, and

(ii) repeated differentiation of each term in f must produce only a finite number of LI terms.

Variation of parameters, on the other hand, merely requires L to be linear. According to the method, we vary the parameters (i.e., the constants of integration in y_h) $C_1, \ldots, C_n$, and seek $y_p(x) = C_1(x)y_1(x) + \cdots + C_n(x)y_n(x)$. Putting that form into the given differential equation gives one condition on the $C_j(x)$'s. That condition is augmented by $n - 1$ additional conditions that are designed to preclude the presence of derivatives of the $C_j(x)$'s that are of order higher than first.

In Section 3.8 we study the *harmonic oscillator*, both damped and undamped, both free and driven. Of special interest are the concepts of *natural frequency* for the undamped case, *critical damping, amplitude- and frequency-response curves, resonance*, and *beats*. This application is of great importance in engineering and science and should be understood thoroughly.

Finally, Section 3.9 is devoted to systems of linear differential equations. We give an existence/uniqueness theorem and show how to solve systems by elimination.

Chapter 4

Power Series Solutions

PREREQUISITES: This chapter presumes a familiarity with the complex plane and the algebra of complex numbers, material which is covered in Section 21.2.

4.1 Introduction

In Chapter 2 we presented a number of methods for obtaining analytical closed form solutions of first-order differential equations, some of which methods could be applied even to nonlinear equations. In Chapter 3 we studied equations of second order and higher, and found them to be more difficult. Restricting our discussion to linear equations, even then we were successful in developing solutions only for the (important) cases of equations with constant coefficients and Cauchy-Euler equations. We also found that we can solve nonconstant-coefficient equations if we can factor the differential operator, but such factorization can be accomplished only in exceptional cases.

In Chapter 4 we continue to restrict our discussion to linear equations, but we now study nonconstant-coefficient equations. That case is so much more difficult than the constant-coefficient case that we do two things: we consider only second-order equations, and we give up on the prospect of finding solutions in closed form and seek solutions in the form of infinite series.

To illustrate the idea that is developed in the subsequent sections, consider the simple example

$$\frac{dy}{dx} + y = 0. \tag{1}$$

To solve by the series method, we seek a solution in the form of a power series expansion about any desired point $x = x_0$, $y(x) = \sum_{n=0}^{\infty} a_n (x - x_0)^n$, where the a_n coefficients are to be determined so that the assumed form satisfies the given differential equation (1), if possible. If we choose $x_0 = 0$ for simplicity, then

$$y(x) = \sum_{0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots, \tag{2a}$$

and

$$\frac{dy}{dx} = \frac{d}{dx}\left(a_0 + a_1 x + a_2 x^2 + \cdots\right) = a_1 + 2a_2 x + 3a_3 x^2 + \cdots . \tag{2b}$$

Putting (2a,b) into (1) gives

$$\left(a_1 + 2a_2 x + 3a_3 x^2 + \cdots\right) + \left(a_0 + a_1 x + a_2 x^2 + \cdots\right) = 0, \tag{3}$$

or, rearranging terms,

$$(a_1 + a_0) + (2a_2 + a_1)\,x + (3a_3 + a_2)\,x^2 + \cdots = 0. \tag{4}$$

If we realize that the right side of (4) is really $0 + 0x + 0x^2 + \cdots$, then, by equating coefficients of like powers of x on both sides of (4), we obtain $a_1 + a_0 = 0$, $2a_2 + a_1 = 0$, $3a_3 + a_2 = 0$, and so on. Thus,

$$\begin{aligned} a_1 &= -a_0, \\ a_2 &= -a_1/2 = -(-a_0)/2 = a_0/2, \\ a_3 &= -a_2/3 = -(a_0/2)/3 = -a_0/6, \end{aligned} \tag{5}$$

and so on, where a_0 remains arbitrary. Thus, we have

$$y(x) = a_0\left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \cdots\right), \tag{6}$$

as the general solution to (1). Here, a_0 is the constant of integration; we could rename it C, for example, if we wish. Thus, we have the solution – not in closed form but as a power series. In this simple example we are able to "sum the series" into closed form, that is, to identify it as the Taylor series of e^{-x}, so that our general solution is really $y(x) = Ce^{-x}$. However, for nonconstant-coefficient differential equations we are generally not so fortunate, and must leave the solution in series form.

As simple as the above steps appear, there are several questions that need to be addressed before we can have confidence in the result given by (6):

(i) In (2b) we differentiated an infinite series term by term. That is, we interchanged the order of the differentiation and the summation and wrote

$$\frac{d}{dx}\sum a_n x^n = \sum \frac{d}{dx}\left(a_n x^n\right). \tag{7}$$

That step looks reasonable, but observe that it amounts to an interchange in the order of two operations, the summation and the differentiation, and it is possible that reversing their order might give different results. For instance, do we get the same results if we put toothpaste on our toothbrush and then brush, or if we brush and then put toothpaste on the brush?

(ii) Re-expressing (3) in the form of (4) is based on a supposition that we can add series term by term:

$$\sum A_n + \sum B_n = \sum (A_n + B_n). \tag{8}$$

Again, that step looks reasonable, but is it necessarily correct?

(iii) Finally, inferring (5) from (4) is based on a supposition that if

$$\sum A_n x^n = \sum B_n x^n \tag{9}$$

for all x in some interval of interest, then it must be true that $A_n = B_n$ for each n. Though reasonable, does it really follow that for the sums to be the same the corresponding individual terms need to be the same?

Thus, there are some technical questions that we need to address, and we do that in the next section. Our approach, in deriving (6), was heuristic, not rigorous, since we did not attend to the issues mentioned above. We can sidestep the several questions of rigor that arose in deriving the series (6) if, instead, we verify, a posteriori, that (6) does satisfy the given differential equation (1). However, that procedure begs exactly the same questions: termwise differentiation of the series, termwise addition of series, and equating the coefficients of like powers of x on both sides of the equation.

Here is a brief outline of this chapter:

4.2 *Power Series Solutions.* In Section 4.2, we review infinite series, power series, and Taylor series, then we show how to find solutions to the equation $y'' + p(x)y' + q(x)y = 0$ in the form of a power series about a chosen point x_0 if $p(x)$ and $q(x)$ are sufficiently well-behaved at x_0.

4.3 *The Method of Frobenius.* If $p(x)$ and $q(x)$ are not sufficiently well-behaved at x_0, then the singular behavior of p and/or q gets passed on in some form to the solutions of the differential equation; hence those solutions cannot be found in power series form. Yet, if $p(x)$ and $q(x)$ are not too singular at x_0, then solutions can still be found, but in a more general form, a so-called Frobenius series. Section 4.3 puts forward the theoretical base for such solutions and the procedure whereby to obtain them.

4.4 *Legendre Functions.* This section focuses on a specific important example, the Legendre equation $(1 - x^2)y'' - 2xy' + \lambda y = 0$, where λ is a constant.

4.5 *Singular Integrals; Gamma Function.* Singular integrals are defined and their convergence is discussed. An important singular integral, the gamma function, is introduced and studied.

4.6 *Bessel Functions.* Besides the Legendre equation, we need to study the extremely important Bessel equation, $x^2y'' + xy' + (x^2 - \nu^2)y = 0$, where ν is a constant, but preparatory to that study we first need to introduce singular integrals and the gamma function, which will be needed again in Chapter 5 in any case.

4.2 Power Series Solutions

4.2.1. Review of power series. Whereas a finite sum,

$$\sum_{k=1}^{N} a_k = a_1 + a_2 + \cdots + a_N, \tag{1}$$

is well-defined thanks to the commutative and associative laws of addition, an infinite sum, or **infinite series**,

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + a_3 + \cdots, \tag{2}$$

is not. For example, is the series $\sum_1^\infty (-1)^{k-1} = 1 - 1 + 1 - 1 + \cdots$ equal to $(1-1) + (1-1) + \cdots = 0 + 0 + \cdots = 0$? Is it (by grouping differently) $1 - (1-1) - (1-1) - \cdots = 1 - 0 - 0 - \cdots = 1$? In fact, besides grouping the numbers in different ways we could rearrange their order as well. The point, then, is that (2) is not self-explanatory, it needs to be defined; we need to decide, or be told, how to do the calculation. To give the traditional definition of (2), we first define the sequence of **partial sums** of the series (2) as

$$s_1 = a_1, \qquad s_2 = a_1 + a_2, \qquad s_3 = a_1 + a_2 + a_3, \tag{3}$$

and so on:

$$s_n = \sum_{k=1}^{n} a_k, \tag{4}$$

where a_k is called the kth **term** of the series. If the limit of the sequence s_n exists, as $n \to \infty$, and equals some number s, then we say that the series (2) is **convergent**, and that it **converges to** s; otherwise it is **divergent**. That is, an infinite series is defined as the limit (if that limit exists) of its sequence of partial sums:

$$\sum_{k=1}^{\infty} a_k \equiv \lim_{n\to\infty} \sum_{k=1}^{n} a_k = \lim_{n\to\infty} s_n = s. \tag{5}$$

That definition, known as **ordinary convergence**, is not the only one possible. For instance, another definition, due to Cesàro, is discussed in the exercises. However, ordinary convergence is the traditional definition and is the one that is understood unless specifically stated otherwise.

Recall from the calculus that by $\lim_{n\to\infty} s_n = s$, in (5), we mean that to each number $\epsilon > 0$, no matter how small, there exists an integer N such that $|s - s_n| < \epsilon$ for all $n > N$. (Logically, the words "no matter how small" are unnecessary, but we include them for emphasis.) In general, the smaller the chosen ϵ, the larger the N that is needed, so that N is a function of ϵ.

The significance of the limit concept cannot be overstated, for in mathematics it is often as limits of "old things" that we introduce "new things." For instance, the derivative is introduced as the limit of a difference quotient, the Riemann integral is introduced as the limit of a sequence of Riemann sums, infinite series are introduced as limits of sequences of partial sums, and so on.

To illustrate the definition of convergence given above, consider two simple examples. The series $1 + 1 + 1 + \cdots$ diverges because $s_n = n$ fails to approach a limit as $n \to \infty$. However, for a series to diverge its partial sums need not grow unboundedly. For instance, the series $1-1+1-1+\cdots$, mentioned above, diverges because its sequence of partial sums (namely, $1, 0, 1, 0, 1, \ldots$) fails to approach a limit. Of course, determining whether a series is convergent or divergent is usually much harder than for these examples. Ideally, one would like a theorem that gives necessary and sufficient conditions for convergence. Here is such a theorem.

THEOREM 4.2.1 *Cauchy Convergence Theorem*
An infinite series is convergent if and only if its sequence of partial sums s_n is a *Cauchy sequence* – that is, if to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| < \epsilon$ for all m and n greater than N.

Unfortunately, this theorem is difficult to apply, so one develops (in the calculus) an array of theorems (i.e., tests for convergence/divergence), each of which is more specialized (and hence less powerful) than the Cauchy convergence theorem, but easier to apply. For instance, if in Theorem 4.2.1 we set $m = n - 1$, then the stated condition becomes: to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| = |a_n| < \epsilon$ for all $n > N$. The latter is equivalent to saying that $a_n \to 0$ as $n \to \infty$. Thus, we have the specialized, but readily applied, theorem that *for the series $\sum^{\infty} a_n$ to converge, it is necessary (but not sufficient) that $a_n \to 0$ as $n \to \infty$.* From this theorem it follows immediately that the series $1 + 1 + 1 + \cdots$ and $1 - 1 + 1 - 1 + \cdots$, cited above, both diverge because in each case the terms do not tend to zero.

Let us now focus on the specific needs of this chapter, **power series** – that is, series of the form

$$\sum_{0}^{n} a_n(x - x_0)^n = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots, \tag{6}$$

where the a_n's are numbers called the **coefficients** of the series, x is a variable, and x_0 is a fixed point called the **center** of the series. We say that the expansion is "about the point x_0." In a later chapter we study complex series, but in this chapter we restrict all variables and constants to be real. Notice that the quantity $(x - x_0)^n$ on the left side of (6) is the indeterminate form 0^0 when $n = 0$ and $x = x_0$; that form must be interpreted as 1 if the leading term of the series is to be a_0, as desired.

The terms in (6) are now functions of x rather than numbers, so that the series may converge at some points on the x axis and diverge at others. At the very least (6) converges at $x = x_0$ since then it reduces to the single term a_0.

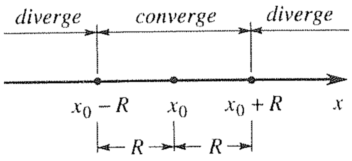


Figure 1. Interval of convergence of power series.

THEOREM 4.2.2 *Interval of Convergence of Power Series*
The power series (6) converges at $x = x_0$. If it converges at other points as well, then those points necessarily comprise an interval $|x - x_0| < R$ centered at x_0 and, possibly, one or both endpoints of that interval (Fig. 1), where R can be determined from either of the formulas

$$R = \frac{1}{\lim\limits_{n\to\infty} \left|\dfrac{a_{n+1}}{a_n}\right|} \qquad \text{or,} \qquad R = \frac{1}{\lim\limits_{n\to\infty} \sqrt[n]{|a_n|}}, \tag{7a,b}$$

if the limits in the denominators exist and are nonzero. If the limits in (7a,b) are zero, then (6) converges for all x (i.e., for every finite x, no matter how large), and we say that "$R = \infty$." If the limits fail to exist by virtue of being infinite, then $R = 0$ and (6) converges only at x_0.

We call $|x - x_0| < R$ the **interval of convergence**, and R the **radius of convergence**. If a power series converges to a function f on some interval, we say that it **represents** f on that interval, and we call f its **sum function**.

EXAMPLE 1. Consider $\sum_0^\infty n!\, x^n$, so $a_n = n!$ and $x_0 = 0$. Then (7a) is easier to apply than (7b), and gives $R = 1/\lim\limits_{n\to\infty} \dfrac{(n+1)!}{n!} = 1/\lim\limits_{n\to\infty}(n+1) = 1/\infty = 0$, so the series converges only at $x = x_0 = 0$. ▮

EXAMPLE 2. Consider $\sum_0^\infty (-1)^n \left[(x+5)/2\right]^n$. Then $a_n = (-1)^n/2^n$, $x_0 = -5$, and (7a) gives $R = 1/\lim\limits_{n\to\infty}\left|\dfrac{(-1)^{n+1}}{2^{n+1}}\dfrac{2^n}{(-1)^n}\right| = 1/\lim\limits_{n\to\infty}\dfrac{1}{2} = 1/\left(\dfrac{1}{2}\right) = 2$, so the series converges in $|x+5| < 2$ and diverges in $|x+5| > 2$. For $|x+5| = 2$ ($x = -7, -3$) the theorem gives no information. However, we see that for $x = -7$ and -3 the terms do not tend to zero as $n \to \infty$, so the series diverges for $x = -7$ and -3. ▮

EXAMPLE 3. Consider $\sum\limits_4^\infty \dfrac{(x-1)^n}{(n+1)^n}$. Then $a_n = (n+1)^{-n}$, $x_0 = 1$, and (7b) gives $R = 1/\lim\limits_{n\to\infty}\sqrt[n]{(n+1)^{-n}} = 1/\lim\limits_{n\to\infty}\dfrac{1}{n+1} = 1/0 = \infty$, so the series converges for all x; that is, the interval of convergence is $|x - 1| < \infty$. ▮

EXAMPLE 4. Consider the series

$$1 + \frac{(x-3)^2}{5} + \frac{(x-3)^4}{5^2} + \cdots = \sum_0^\infty \frac{1}{5^n}(x-3)^{2n}. \tag{8}$$

This series is not of the form (6) because the powers of $x-3$ proceed in steps of 2. However, if we set $X = (x-3)^2$, then we have the standard form $\sum_0^\infty \frac{1}{5^n} X^n$, with $a_n = 1/5^n$ and $\lim_{n\to\infty} \left| \frac{a_{n+1}}{a_n} \right| = \lim_{n\to\infty} \left| \frac{5^n}{5^{n+1}} \right| = \frac{1}{5}$. Thus, $R = 5$, and the series converges in $|X| < 5$ (i.e., $|x-3| < \sqrt{5}$), and diverges in $|X| > 5$ (i.e., $|x-3| > \sqrt{5}$). ∎

Recall from our introductory example, in Section 4.1, that several questions arose regarding the manipulation of power series. The following theorem answers those questions and, therefore, will be needed when we apply the power series method of solution.

THEOREM 4.2.3 *Manipulation of Power Series*
(a) *Termwise differentiation (or integration) permissible.* A power series may be differentiated (or integrated) termwise (i.e., term by term) within its interval of convergence I. The series that results has the same interval of convergence I and represents the derivative (or integral) of the sum function of the original series.
(b) *Termwise addition (or subtraction or multiplication) permissible.* Two power series (about the same point x_0) may be added (or subtracted or multiplied) termwise within their common interval of convergence I. The series that results has the same interval of convergence I and represents the sum (or difference or product) of their two sum functions.
(c) *If two power series are equal, then their corresponding coefficients must be equal.* That is, for

$$\sum_0^\infty a_n(x-x_0)^n = \sum_0^\infty b_n(x-x_0)^n \tag{9}$$

to hold in some common interval of convergence, it must be true that $a_n = b_n$ for each n. In particular, if

$$\sum_0^\infty a_n(x-x_0)^n = 0 \tag{10}$$

in some interval, then each a_n must be zero.

Part (a) means that if $f(x) = \sum_0^\infty a_n(x-x_0)^n$ within I, then

$$f'(x) = \frac{d}{dx}\sum_0^\infty a_n(x-x_0)^n = \sum_0^\infty \frac{d}{dx}\left[a_n(x-x_0)^n\right] = \sum_1^\infty n a_n (x-x_0)^{n-1} \tag{11}$$

and

$$\begin{aligned}\int_a^b f(x)\,dx &= \int_a^b \sum_0^\infty a_n(x-x_0)^n\,dx \\ &= \sum_0^\infty a_n \int_a^b (x-x_0)^n\,dx \\ &= \sum_0^\infty a_n \frac{(b-x_0)^{n+1}-(a-x_0)^{n+1}}{n+1} \qquad (12)\end{aligned}$$

within I, where a, b are any two points within I.

Part (b) means that if $f(x) = \sum_0^\infty a_n(x-x_0)^n$ and $g(x) = \sum_0^\infty b_n(x-x_0)^n$ on I, then

$$f(x) \pm g(x) = \sum_0^\infty (a_n \pm b_n)(x-x_0)^n, \qquad (13)$$

and, with $z = x - x_0$ for brevity,

$$\begin{aligned}f(x)g(x) &= \left(\sum_0^\infty a_n z^n\right)\left(\sum_0^\infty b_n z^n\right) \\ &= (a_0 + a_1 z + \cdots)(b_0 + b_1 z + \cdots) \\ &= a_0\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) + a_1 z\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) \\ &\quad + a_2 z^2\left(b_0 + b_1 z + b_2 z^2 + \cdots\right) + \cdots \\ &= a_0 b_0 + (a_0 b_1 + a_1 b_0)\, z + \cdots \\ &= \sum_0^\infty (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0)\, z^n \qquad (14)\end{aligned}$$

within I. The series on the right-hand side of (14) is known as the **Cauchy product** of the two series. Of course, if the two convergence intervals have different radii, then the common interval means the smaller of the two.

In summary, we see that convergent power series can be manipulated in essentially the same way as if they were finite-degree polynomials.

The last items to address, before coming to the power series method of solution of differential equations, are Taylor series and analyticity. Recall from the calculus that the **Taylor series** of a given function $f(x)$ about a chosen point x_0, which we denote here as TS $f|_{x_0}$, is defined as the infinite series

$$\begin{aligned}\text{TS } f|_{x_0} &= f(x_0) + \frac{f'(x_0)}{1!}(x-x_0) + \frac{f''(x_0)}{2!}(x-x_0)^2 + \cdots \\ &= \sum_0^\infty \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n, \qquad (15)\end{aligned}$$

where $0! = 1$. The purpose of Taylor series is to represent the given function, so the fundamental question is: does it? Does the Taylor series really converge to $f(x)$ on some x interval, in which case we can write, in place of (15),

$$f(x) = \sum_{0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n. \tag{16}$$

For that to be the case we need three conditions to be met:

(i) First, we need f to *have* a Taylor series (15) about that point. Namely, f must be infinitely differentiable at x_0 so that all of the coefficients $f^{(n)}(x_0)/n!$ in (15) exist.

(ii) Second, we need the resulting series in (15) to *converge* in some interval $|x - x_0| < R$, for $R > 0$.

(iii) Third, we need the sum of the Taylor series to equal f in the interval, so that the Taylor series *represents* f over that interval – which is, after all, our objective.

The third condition might seem strange, for how could the Taylor series of $f(x)$ converge, but to something other than $f(x)$? Such cases can indeed be put forward, but they are somewhat pathological and not likely to be encountered in applications.

If a function is represented in some nonzero interval $|x - x_0| < R$ by its Taylor series [i.e., TS $f|_{x_0}$ exists, and converges to $f(x)$ there], then f is said to be **analytic** at x_0. If a function is not analytic at x_0, then it is **singular** there.

Most functions encountered in applications are analytic for all x, or for all x with the exception of one or more points called **singular points** of f. (Of course, the points are not singular, the function is.) For instance, polynomial functions, $\sin x$, $\cos x$, e^x, and e^{-x} are analytic for all x. On the other hand, $f(x) = 1/(x-1)$ is analytic for all x except $x = 1$, where f and all of its derivatives are undefined, fail to exist. The function $f(x) = \tan x = \sin x/\cos x$ is analytic for all x except $x = n\pi/2$ ($n = \pm 1, \pm 3, \ldots$), where it is undefined because $\cos x$ vanishes in the denominator.

The function $f(x) = x^{4/3}$ is analytic for all x except $x = 0$, for even though $f(0)$ and $f'(0)$ exist, the subsequent derivatives $f''(0)$, $f'''(0)$,... do not (Fig. 2). In fact, $f(x) = x^\alpha$ is singular at $x = 0$ for *any* noninteger value of α.

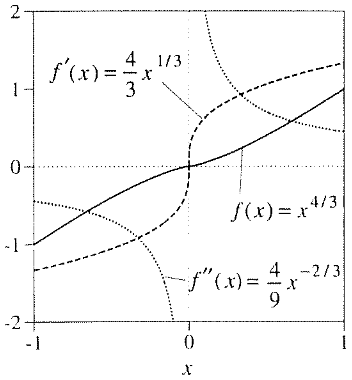


Figure 2. $f(x) = x^{4/3}$ and its first two derivatives.

Observe that there is a subtle difficulty here. We know how to test a given Taylor series for convergence since a Taylor series is a power series, and Theorem 4.2.2 on power series convergence even gives formulas for determining the radius of convergence R. But how can we determine if the sum function (i.e., the function to which the series converges) is the same as the original function f? We won't be able to answer this question until we study complex variable theory, in later chapters. However, we repeat that the cases where the Taylor series of f converges, but not to f, are exceptional and will not occur in the present chapter, so it will suffice to understand analyticity at x_0 to correspond to the convergence of the Taylor series in some nonzero interval about x_0. In fact, it is also exceptional for f to have a Taylor series about a point (i.e., be infinitely differentiable at that point) and to have that Taylor series fail to converge in some nonzero interval about x_0. Thus, *as a rule of thumb that will suffice until we study complex variable theory, we will test a function for analyticity at a given point simply by seeing if it is infinitely differentiable at that point.*

4.2.2. Power series solution of differential equations. We can now state the following basic theorem.

THEOREM 4.2.4 *Power series solution*
If p and q are analytic at x_0, then every solution of

$$y'' + p(x)y' + q(x)y = 0 \tag{17}$$

is too, and can therefore be found in the form

$$y(x) = \sum_0^\infty a_n(x - x_0)^n. \tag{18}$$

Further, the radius of convergence of every solution (18) is at least as large as the smaller of the radii of convergence of TS $p|_{x_0}$ and TS $q|_{x_0}$.

Although we will not prove this theorem, we shall explain why one can expect it to be true. Since p and q are analytic at the chosen point x_0, they admit convergent Taylor series about x_0, so that we can write (17) as

$$y'' + \left[p(x_0) + p'(x_0)(x - x_0) + \cdots\right] y' + \left[q(x_0) + q'(x_0)(x - x_0) + \cdots\right] y = 0. \tag{19}$$

Locally, near x_0, we can approximate (19) as

$$y'' + p(x_0)y' + q(x_0)y = 0,$$

all solutions of which are either exponential or x times an exponential, and are therefore analytic and representable in the form (18), as claimed.

In many applications, $p(x)$ and $q(x)$ are **rational functions**, that is, one polynomial in x divided by another. Let $F(x) = N(x)/D(x)$ be any rational function, where the numerator and denominator polynomials are $N(x)$ and $D(x)$, respectively, and where any common factors have been canceled. It will be shown, when we study complex variable theory, that $F(x)$ is singular only at those points in the complex plane where $D = 0$, at the *zeros* of D, so that a Taylor expansion of F about a point x_0 on the x axis will have a radius of convergence which, we know in advance, will be equal to the distance from x_0 on the x axis to the nearest zero of D in the complex plane. For instance, if $F(x) = (2+3x)/[(4+x)(9+x^2)]$, then D has zeros at -4 and $\pm 3i$. Thus, if we expand F about $x = 2$, say, then the radius of convergence will be the distance from 2 to the nearest zero, which is $+3i$ (or, equally, $-3i$), namely, $\sqrt{13}$ (Fig. 3). If, instead, we expand about $x = -6$, say, then the radius of convergence will be 2, the distance from -6 to the zero of D at -4.

EXAMPLE 5. Solve

$$y'' + y = 0 \tag{20}$$

by the power series method. Of course, this equation is elementary. We know the solution and do not need the power series method to find it. Let us use it nevertheless, as a first example, to illustrate the method.

We can choose the point of expansion x_0 in (18) as any point at which both $p(x)$ and $q(x)$ in (17) are analytic. In the present example, $p(x) = 0$ and $q(x) = 1$ are analytic for all x, so we can choose the point of expansion x_0 to be whatever we like. Let $x_0 = 0$, for simplicity. Then Theorem 4.2.4 assures us that all solutions can be found in the form

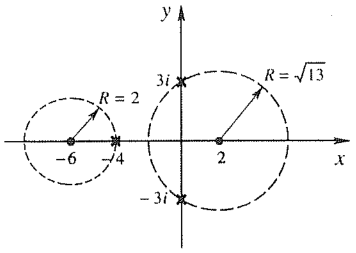


Figure 3. Disks of convergence in z plane ($z = x + iy$).

$$y(x) = \sum_0^\infty a_n x^n, \tag{21a}$$

and their series will have infinite radii of convergence. Within that (infinite) interval of convergence we have, from Theorem 4.2.3(a),

$$y'(x) = \sum_1^\infty n a_n x^{n-1}, \tag{21b}$$

$$y''(x) = \sum_2^\infty n(n-1) a_n x^{n-2}, \tag{21c}$$

so (20) becomes

$$\sum_{n=2}^\infty n(n-1) a_n x^{n-2} + \sum_{n=0}^\infty a_n x^n = 0. \tag{22}$$

We would like to combine the two sums, but to do that we need the exponents of x to be the same, whereas they are $n-2$ and n. To have the same exponents, let us set $n-2=m$ in the first sum, just as one might make a change of variables in an integral. In the integral, one would then need to change the integration limits, if necessary, consistent with the change of variables; here, we need to do likewise with the summation limits. With $n-2=m$, $n=\infty$ corresponds to $m=\infty$, and $n=2$ corresponds to $m=0$, so (22) becomes

$$\sum_{m=0}^\infty (m+2)(m+1) a_{m+2} x^m + \sum_{n=0}^\infty a_n x^n = 0. \tag{23}$$

Next, and we shall explain this step in a moment, let $m = n$ in the first sum in (23):

$$\sum_{n=0}^\infty (n+2)(n+1) a_{n+2} x^n + \sum_{n=0}^\infty a_n x^n = 0 \tag{24}$$

or, with the help of Theorem 4.2.3(b),

$$\sum_{n=0}^{\infty}\left[(n+2)(n+1)a_{n+2}+a_n\right]x^n=0. \tag{25}$$

Finally, it follows from Theorem 4.2.3(c) that each coefficient in (25) must be zero:

$$(n+2)(n+1)a_{n+2}+a_n=0. \qquad (n=0,1,2,\dots) \tag{26}$$

Before using (26), let us explain our setting $m=n$ in (23) since that step might seem to contradict the preceding change of variables $n-2=m$. The point to appreciate is that m in the first sum in (23) is a *dummy index* just as t is a *dummy variable* in $\int_0^1 t^2\,dt$. (We shall use the word *index* for a discrete variable; m takes on only integer values, not a continuous range of values.) Just as $\int_0^1 t^2\,dt=\int_0^1 r^2\,dr=\int_0^1 x^2\,dx=\cdots=\frac{1}{3}$, the sums in (23) are insensitive to whether the dummy index is m or n:

$$\sum_{m=0}^{\infty}(m+2)(m+1)a_{m+2}x^m=2a_2+6a_3x+12a_4x^2+\cdots,$$

and

$$\sum_{n=0}^{\infty}(n+2)(n+1)a_{n+2}x^n=2a_2+6a_3x+12a_4x^2+\cdots$$

are identical, even though the summation indices are different.

Equation (26) is known as a **recursion** (or **recurrence**) **formula** on the unknown coefficients since it gives us the nth coefficient in terms of preceding ones. Specifically,

$$a_{n+2}=-\frac{1}{(n+2)(n+1)}a_n, \qquad (n=0,1,2,\dots) \tag{27}$$

so that

$$\begin{aligned}
n=0: &\quad a_2=-\frac{1}{(2)(1)}a_0,\\
n=1: &\quad a_3=-\frac{1}{(3)(2)}a_1,\\
n=2: &\quad a_4=-\frac{1}{(4)(3)}a_2=\frac{1}{(4)(3)(2)(1)}a_0=\frac{1}{4!}a_0,\\
n=3: &\quad a_5=-\frac{1}{(5)(4)}a_3=\frac{1}{5!}a_1,
\end{aligned} \tag{28}$$

and so on, where a_0 and a_1 remain arbitrary and are, in effect, the integration constants. Putting these results into (21a) gives

$$\begin{aligned}
y(x)&=a_0+a_1x-\frac{1}{2!}a_0x^2-\frac{1}{3!}a_1x^3+\frac{1}{4!}a_0x^4+\frac{1}{5!}a_1x^5+\cdots\\
&=a_0\left(1-\frac{1}{2!}x^2+\frac{1}{4!}x^4-\cdots\right)+a_1\left(x-\frac{1}{3!}x^3+\frac{1}{5!}x^4-\cdots\right)
\end{aligned} \tag{29}$$

or

$$y(x)=a_0y_1(x)+a_1y_2(x), \tag{30}$$

where $y_1(x)$ and $y_2(x)$ are the series within the first and second pairs of parentheses, respectively. From their series, we recognize $y_1(x)$ as $\cos x$ and $y_2(x)$ as $\sin x$ but, in general, we can't expect to identify the power series in terms of elementary functions because normally we reserve the power series method for nonelementary equations (except for pedagogical examples such as this). Thus, let us continue to call the series "$y_1(x)$" and "$y_2(x)$."

We don't need to check the series for convergence because Theorem 4.2.4 guarantees that they will converge for all x. We should, however, check to see if y_1, y_2 are LI (linearly independent), so that (30) is a general solution of (20). To do so, it suffices to evaluate the Wronskian $W[y_1, y_2](x)$ at a single point, say $x = 0$:

$$W[y_1,y_2](x) = \begin{vmatrix} y_1(0) & y_2(0) \\ y_1'(0) & y_2'(0) \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1, \tag{31}$$

which is nonzero. It follows from that result and Theorem 3.2.3 that y_1, y_2 are LI on the entire x axis, so (30) is indeed a general solution of (20) for all x. Actually, since there are only two functions it would have been easier to apply Theorem 3.2.4: y_1, y_2 are LI because neither one is a constant multiple of the other.

COMMENT 1. To evaluate $y_1(x)$ or $y_2(x)$ at a given x, we need to add enough terms of the series to achieve the desired accuracy. For small values of x (i.e., for x's that are close to the point of expansion x_0, which in this case is 0) just a few terms may suffice. For example, the first four terms of $y_1(x)$ give $y_1(0.5) = 0.877582$, whereas the exact value is 0.877583 (to six decimal places). As x increases, more and more terms are needed for comparable accuracy. The situation is depicted graphically in Fig. 4, where we plot the partial sums s_3 and s_6, along with the sum function $y_1(x)$ (i.e., $\cos x$). Observe that the larger n is, the broader is the x interval over which the n-term approximation s_n stays close to the sum function. However, whereas $y_1(x)$ is oscillatory and remains between -1 and $+1$, $s_n(x)$ is a polynomial, and therefore it eventually tends to $+\infty$ or $-\infty$ as x increases ($-\infty$ if n is even and $+\infty$ if n is odd). Observe that if we do need to add a great many terms, then it is useful to have an expression for the general term in the series. In this example it is not hard to establish that

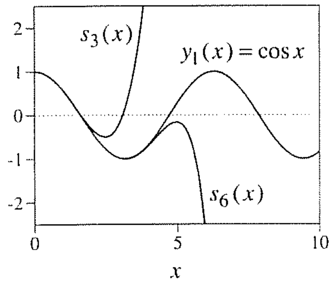


Figure 4. Partial sums of $y_1(x)$, compared with $y_1(x)$.

$$y_1(x) = \sum_0^\infty (-1)^n \frac{x^{2n}}{(2n)!}, \qquad y_2(x) = \sum_0^\infty (-1)^n \frac{x^{2n+1}}{(2n+1)!}. \tag{32}$$

COMMENT 2. Besides obtaining the values of the solutions $y_1(x)$ and $y_2(x)$, one is usually interested in determining some of their properties. Some properties can be obtained directly from the series. For instance, in this case we can see that $y_1(-x) = y_1(x)$ and $y_2(-x) = -y_2(x)$ (so that the graphs of y_1 and y_2 are symmetric and antisymmetric, respectively, about $x = 0$), and that $y_1'(x) = -y_2(x)$ and $y_2'(x) = y_1(x)$. The differential equation is also a source of information.

COMMENT 3. We posed (20) without initial conditions. Suppose that we wish to impose the initial conditions $y(0) = 4$ and $y'(0) = -1$. Then, from (30),

$$\begin{aligned} y(0) &= 4 = a_0 y_1(0) + a_1 y_2(0), \\ y'(0) &= -1 = a_0 y_1'(0) + a_1 y_2'(0). \end{aligned} \tag{33}$$

From the series representations of y_1 and y_2 in (30), we see that $y_1(0) = 1$, $y_2(0) = 0$, $y_1'(0) = 0$, and $y_2'(0) = 1$, so we can solve (33) for a_0 and a_1: $a_0 = 4$ and $a_1 = -1$, hence the desired particular solution is $y(x) = 4y_1(x) - y_2(x)$, on $-\infty < x < \infty$. ▮

We can now see more clearly how to select the point of expansion x_0, besides selecting it to be a point at which p and q in (17) are analytic. We have emphasized that the series solutions are especially convenient when the calculation point x is close to x_0, for then only a few terms of the series may suffice for the desired accuracy. Thus, if our interest is limited to the interval $6 < x < 10$, say, then it would make sense to seek series solutions about a point somewhere in that interval, such as a midpoint or endpoint, rather than about some distant point such as $x = 0$.

In the event that initial conditions are supplied at some point x_i, then it is especially helpful to let x_0 be x_i because when we apply the initial conditions we will need to know the values of $y_1(x_i)$, $y_2(x_i)$, $y_1'(x_i)$, and $y_2'(x_i)$, as we did in (33). If x_0 is other than x_i, then each of these evaluations requires the summing of an infinite series, whereas if it is chosen as x_i then these evaluations are trivial (as in Comment 3, above).

EXAMPLE 6. Solve the initial-value problem

$$(x-1)y'' + y' + 2(x-1)y = 0, \qquad y(4) = 5, \quad y'(4) = 0 \tag{34}$$

on the interval $4 \le x < \infty$. To get (34) into the standard form $y'' + p(x)y' + q(x)y = 0$, we divide by $x - 1$ (which is permissible since $x - 1 \neq 0$ on the interval of interest):

$$y'' + \frac{1}{x-1}y' + 2y = 0, \tag{35}$$

so $p(x) = 1/(x-1)$ and $q(x) = 2$. These are analytic for all x except $x = 1$, where $p(x)$ is undefined. In particular, they are analytic at the initial point $x = 4$, so let us choose $x_0 = 4$ and seek

$$y(x) = \sum_0^\infty a_n(x-4)^n. \tag{36}$$

To proceed we can use either form (34) or (35). Since we are expanding each term in the differential equation about $x = 4$, we need to expand $x - 1$ and $2(x-1)$ if we use (34), or the $1/(x-1)$ factor if we use (35). The former is easier since

$$x - 1 = 3 + (x-4) \tag{37}$$

is merely a two-term Taylor series, whereas (Exercise 6)

$$\frac{1}{x-1} = \frac{1}{3}\sum_0^\infty \frac{(-1)^n}{3^n}(x-4)^n \tag{38}$$

is an infinite series. Thus, let us use (34). Putting (36) and its derivatives and (37) into (34) gives

$$[3 + (x-4)]\sum_2^\infty n(n-1)a_n(x-4)^{n-2} + \sum_1^\infty na_n(x-4)^{n-1}$$

$$+2\left[3+(x-4)\right]\sum_{0}^{\infty}a_n(x-4)^n=0 \tag{39}$$

or, absorbing the $3+(x-4)$ terms into the series that they multiply and setting $z=x-4$ for compactness,

$$\sum_{2}^{\infty}3n(n-1)a_nz^{n-2}+\sum_{2}^{\infty}n(n-1)a_nz^{n-1}$$
$$+\sum_{1}^{\infty}na_nz^{n-1}+\sum_{0}^{\infty}6a_nz^n+\sum_{0}^{\infty}2a_nz^{n+1}=0. \tag{40}$$

To adjust all z exponents to n, let $n-2=m$ in the first sum, $n-1=m$ in the second and third, and $n+1=m$ in the last:

$$\sum_{0}^{\infty}3(m+2)(m+1)a_{m+2}z^m+\sum_{1}^{\infty}(m+1)ma_{m+1}z^m$$
$$+\sum_{0}^{\infty}(m+1)a_{m+1}z^m+\sum_{0}^{\infty}6a_nz^n+\sum_{1}^{\infty}2a_{m-1}z^m=0. \tag{41}$$

Next, we change all of the m indices to n. Then we have z^n in each sum, but we cannot yet combine the five sums because the lower summation limits are not all the same; three are 0 and two are 1. We can handle that problem as follows. The lower limit in the second sum can simply be changed from 1 to 0 because the zeroth term is zero anyhow (due to the m factor). And the lower limit in the last sum can be changed to 0 if we simply agree that the a_{-1}, that occurs in the zeroth term, be zero by definition. Then (41) becomes

$$\sum_{0}^{\infty}3(n+2)(n+1)a_{n+2}z^n+\sum_{0}^{\infty}(n+1)na_{n+1}z^n$$
$$+\sum_{0}^{\infty}(n+1)a_{n+1}z^n+\sum_{0}^{\infty}6a_nz^n+\sum_{0}^{\infty}2a_{n-1}z^n=0 \tag{42}$$

or

$$\sum_{0}^{\infty}\left[3(n+2)(n+1)a_{n+2}+(n+1)^2a_{n+1}+6a_n+2a_{n-1}\right]z^n=0, \tag{43}$$

with $a_{-1}\equiv 0$.

Setting the square-bracketed coefficients to zero for each n [Theorem 4.2.3(c)] then gives the recursion formula

$$3(n+2)(n+1)a_{n+2}+(n+1)^2a_{n+1}+6a_n+2a_{n-1}=0$$

or

$$a_{n+2}=-\frac{n+1}{3(n+2)}a_{n+1}-\frac{2}{(n+2)(n+1)}a_n-\frac{2}{3(n+2)(n+1)}a_{n-1} \tag{44}$$

for $n = 0, 1, 2, \ldots$. Thus,

$$\begin{aligned}
n = 0: \quad & a_2 = -\frac{1}{6}a_1 - a_0 - \frac{1}{3}a_{-1} = -\frac{1}{6}a_1 - a_0 \\
n = 1: \quad & a_3 = -\frac{2}{9}a_2 - \frac{1}{3}a_1 - \frac{1}{9}a_0 \\
& \quad = -\frac{2}{9}\left(-\frac{1}{6}a_1 - a_0\right) - \frac{1}{3}a_1 - \frac{1}{9}a_0 = -\frac{8}{27}a_1 + \frac{1}{9}a_0 \\
n = 2: \quad & a_4 = -\frac{1}{4}a_3 - \frac{1}{6}a_2 - \frac{1}{18}a_1 \\
& \quad = -\frac{1}{4}\left(-\frac{8}{27}a_1 + \frac{1}{9}a_0\right) - \frac{1}{6}\left(-\frac{1}{6}a_1 - a_0\right) - \frac{1}{18}a_1 \\
& \quad = \frac{5}{108}a_1 + \frac{5}{36}a_0,
\end{aligned} \tag{45}$$

and so on, where a_0 and a_1 remain arbitrary. Putting these expressions for the a_n's back into (36) then gives

$$\begin{aligned}
y(x) &= a_0 + a_1(x-4) + \left(-\frac{1}{6}a_1 - a_0\right)(x-4)^2 + \left(-\frac{8}{27}a_1 + \frac{1}{9}a_0\right)(x-4)^3 \\
&\quad + \left(\frac{5}{108}a_1 + \frac{5}{36}a_0\right)(x-4)^4 + \cdots \\
&= a_0\left[1 - (x-4)^2 + \frac{1}{9}(x-4)^3 + \frac{5}{36}(x-4)^4 + \cdots\right] \\
&\quad + a_1\left[(x-4) - \frac{1}{6}(x-4)^2 - \frac{8}{27}(x-4)^3 + \frac{5}{108}(x-4)^4 + \cdots\right] \\
&= a_0 y_1(x) + a_1 y_2(x),
\end{aligned} \tag{46}$$

where $y_1(x), y_2(x)$ are the functions represented by the bracketed series. To test y_1, y_2 for linear independence it is simplest to use Theorem 3.2.4: y_1, y_2 are LI because neither one is a constant multiple of the other. Thus, $y(x) = a_0 y_1(x) + a_1 y_2(x)$ is a general solution of $(x-1)y'' + y' + 2(x-1)y = 0$.

Imposing the initial conditions is easy because the expansions are about the initial point $x = 4$:

$$\begin{aligned}
y(4) &= 5 = a_0 y_1(4) + a_1 y_2(4) = a_0(1) + a_1(0), \\
y'(4) &= 0 = a_0 y_1'(4) + a_1 y_2'(4) = a_0(0) + a_1(1),
\end{aligned} \tag{47}$$

so $a_0 = 5$ and $a_1 = 0$, and

$$y(x) = 5y_1(x) = 5\left[1 - (x-4)^2 + \frac{1}{9}(x-4)^3 + \frac{5}{36}(x-4)^4 + \cdots\right] \tag{48}$$

is the desired particular solution.

COMMENT. Recall that Theorem 4.2.4 guaranteed that the power series solution would have a radius of convergence R at least as large as 3 – namely, the distance from the center of the expansion ($x_0 = 4$) to the singularity in $1/(x-1)$ at $x = 1$. For comparison, let us determine R from our results. In this example it is difficult to obtain a general expression for a_n. (Indeed, we didn't attempt to; we were content to develop the first several terms of the series, knowing that we could obtain as many more as we wish, from the recursion formula.) Can we obtain R without an explicit expression for a_n? Yes, we can use the recursion formula (44), which tells us that $a_{n+2} \sim -\frac{1}{3}a_{n+1}$ as $n \to \infty$ or, equivalently, $a_{n+1} \sim -\frac{1}{3}a_n$. Then, from (7a),

$$R = \frac{1}{\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|} = \frac{1}{\lim_{n\to\infty}\left|-\frac{1}{3}\right|} = \frac{1}{\frac{1}{3}} = 3.$$

Thus, if we were hoping to obtain the solution over the entire interval $4 \le x < \infty$ we are disappointed to find that the power series converges only over $1 < x < 7$, and hence only over the $4 \le x < 7$ part of the problem domain. Does this result mean that the solution is singular at $x = 7$ and can't be continued beyond, or that it doesn't exist beyond $x = 7$? No, the convergence is simply being limited by the singularity at $x = 1$, which lies outside of the problem domain $4 \le x < \infty$. For further discussion of this point, see Exercise 12. ▮

Closure. In Section 4.2.1 we reviewed the basic concepts of series and power series and, in Theorem 4.2.3, we listed the properties of power series that are needed to solve differential equations. In Section 4.2.2 we provided a basis for the power series solution method of Theorem 4.2.4 and then showed, by two examples, how to implement it.

It is best to use summation notation, as we did in Examples 5 and 6, because it is more concise and leads to the recursion relation. (But that notation is not essential to the method; for example, we did not use it in our introductory example in Section 4.1.) The recursion relation is important because it permits the calulation of as many coefficients of the series as we desire, and because it can be used in determining the radius of convergence of the resulting series solutions.

The method may be outlined as follows:

(1) Write the differential equation in the standard form $y'' + p(x)y' + q(x)y = 0$ to identify $p(x)$ and $q(x)$ and their singularities (if any).

(2) Choose an expansion point x_0 at which p and q are analytic. If initial conditions are given at some point, it is suggested that that point be used as x_0.

(3) The least possible radius of convergence can be predicted as the distance (in the complex plane) from x_0 to the nearest singular point of p and q.

(4) Seeking $y(x)$ in the form of a power series about x_0, put that form into the differential equation, and also expand all coefficients of y'', y', y about x_0 as well.

(5) By changing dummy indices of summation and lower summation limits, as necessary, obtain a form where each summation has the same exponent on $x - x_0$ and the same summation limits.

(6) Combine the sums into a single sum.

(7) Set the coefficient of $(x - x_0)^n$, within that sum, to zero; that step gives the recursion formula.

(8) From the recursion formula, obtain as many of the coefficients as desired and hence the solution form $y(x) = Ay_1(x) + By_2(x)$, where A, B are arbitrary constants and $y_1(x), y_2(x)$ are power series. If possible, obtain expressions for the general term in each of those series.

(9) Verify that y_1, y_2 are LI.

Computer software. One can use computer software to generate Taylor series and also to obtain power series solutions to differential equations. Using *Maple*, for instance, the relevant commands are **taylor** and **dsolve**.

For example, to obtain the Taylor series of $1/(a - x)$ about $x = 0$, up to terms of third order, where a is a constant, enter

$$\text{taylor}(1/(a-x),\ x = 0,\ 4);$$

and return. The result is

$$\frac{1}{a} + \frac{1}{a^2}x + \frac{1}{a^3}x^2 + \frac{1}{a^4}x^3 + O(x^4)$$

where the $O(x^4)$ denotes that there are more terms, of order 4 and higher.

To obtain a power series solution to $y'' + y = 0$ about the point $x = 0$, enter

$$\text{dsolve}(\text{diff}(y(x),\ x, x) + y(x) = 0,\ y(x),\ \text{type} = \text{series});$$

and return. The result is

$$y(x) = y(0) + D(y)(0)x - \frac{1}{2}y(0)x^2 - \frac{1}{6}D(y)(0)x^3 + \frac{1}{24}y(0)x^4$$
$$+\frac{1}{120}D(y)(0)x^5 + O(x^6)$$

where $D(y)(0)$ means $y'(0)$. The default is to expand about $x = 0$ and to go as far as the fifth-order term. If we want an expansion about $x = 4$, say, and through the seventh-order term, enter

$$\text{Order} := 8;$$

to set the order then return and enter

$$\text{dsolve}(\{\text{diff}(y(x),\ x, x) + y(x) = 0,\ y(4) = a,\ D(y)(4) = b\},$$
$$y(x),\ \text{type} = \text{series});$$

and return. Here, the initial conditions merely serve to establish the point about which the expansion is desired. The result is

$$y(x) = a + b(x-4) - \frac{1}{2}a(x-4)^2 - \frac{1}{6}b(x-4)^3 + \frac{1}{24}a(x-4)^4$$
$$+\frac{1}{120}b(x-4)^5 - \frac{1}{720}a(x-4)^6 - \frac{1}{5040}b(x-4)^7 + O\left((x-4)^8\right)$$

EXERCISES 4.2

1. Use (7a) or (7b) to determine the radius of convergence R of the given power series.

(a) $\displaystyle\sum_0^\infty nx^n$ (b) $\displaystyle\sum_0^\infty (-1)^n n^{1000} x^n$

(c) $\displaystyle\sum_5^\infty e^n x^n$ (d) $\displaystyle\sum_0^{1000} n!\, x^n$

(e) $\displaystyle\sum_0^\infty \left(\frac{x+3}{2}\right)^n$ (f) $\displaystyle\sum_2^\infty (n-1)^3(x-5)^n$

(g) $\displaystyle\sum_1^\infty \frac{n^{50}}{n!}(x+7)^n$ (h) $\displaystyle\sum_2^\infty (\ln n)^{n+1}(x-2)^n$

(i) $\displaystyle\sum_3^\infty \frac{(-1)^n}{4^n}(x+2)^{3n}$ (j) $\displaystyle\sum_0^\infty \frac{n}{2^n}(x-5)^{2n}$

(k) $\displaystyle\sum_0^\infty \frac{n^6}{3^n+n}(x+4)^{8n+1}$ (l) $\displaystyle\sum_0^\infty \frac{(-1)^{n+1}}{4^n+1}(x-3)^{2n+1}$

2. Determine the radius of convergence R of the Taylor series expansion of the given rational function, about the specified point x_0, using the ideas given in the paragraph preceding Example 5. Also, prepare a sketch analogous to those in Fig. 3.

(a) $\dfrac{1}{x^2+1}, \quad x_0 = 0$

(b) $\dfrac{1}{x^2+9}, \quad x_0 = 2$

(c) $\dfrac{x^3-2x+1}{x+7}, \quad x_0 = 5$

(d) $\dfrac{x+1}{x+2}, \quad x_0 = -26$

(e) $\dfrac{(x+1)^5}{x^2+3x+2}, \quad x_0 = -4$

(f) $\dfrac{x^2-3x+1}{x^2+2x+4}, \quad x_0 = 3$

(g) $\dfrac{x^2+x-2}{x^3-x^2+4x-4}, \quad x_0 = 2$

(h) $\dfrac{x^2-3x+2}{x-1}, \quad x_0 = 0$

3. Work out the Taylor series of the given function, about the given point x_0, and use (7a) or (7b) to determine its radius of convergence.

(a) $e^x, \quad x_0 = 1$ (b) $e^{-x}, \quad x_0 = -2$
(c) $\sin x, \quad x_0 = \pi$ (d) $\sin x, \quad x_0 = \pi/2$
(e) $\cos x, \quad x_0 = \pi/2$ (f) $\cos x, \quad x_0 = \pi$
(g) $\cos x, \quad x_0 =5$ (h) $\ln x, \quad x_0 = 1$
(i) $x^2, \quad x_0 = 3$ (j) $2x^3 - 4; \quad x_0 = 0$
(k) $\cos(x-2), \quad x_0 = 2$ (l) $\dfrac{1}{1-x^{10}}, \quad x_0 = 0$
(m) $\dfrac{x^2}{1+x^{18}}, \quad x_0 = 0$ (n) $\sin(3x^{10}), \quad x_0 = 0$

4. Use computer software to obtain the first 12 nonzero terms in the Taylor series expansion of the given function f, about the given point x_0, and obtain a computer plot of f and the partial sums $s_3(x)$, $s_6(x)$, $s_9(x)$, and $s_{12}(x)$ over the given interval I.

(a) $f(x) = e^{-x}, \quad x_0 = 0, \quad I:\ 0 < x < 4$
(b) $f(x) = \sin x, \quad x_0 = 0, \quad I:\ 0 < x < 10$
(c) $f(x) = \ln x, \quad x_0 = 1, \quad I:\ 0 < x < 2$
(d) $f(x) = 1/(1-x), \quad x_0 = 0, \quad I:\ -1 < x < 1$
(e) $f(x) = 1/x, \quad x_0 = 2, \quad I:\ 0 < x < 4$
(f) $f(x) = 1/(1+x^2), \quad x_0 = 0, \quad I:\ -1 < x < 1$
(g) $f(x) = 4/(4+x+x^2), \quad x_0 = 0, \quad I:\ -1.3 < x < 0.36$

5. (*Geometric series*) (a) Show that

$$\boxed{\frac{1}{1-x} = 1 + x + x^2 + \cdots + x^{n-1} + \frac{x^n}{1-x}} \qquad (5.1)$$

is an *identity* for all $x \neq 1$ and any positive integer n, by multiplying through by $1-x$ (which is nonzero since $x \neq 1$) and simplifying.

(b) The identity (5.1) can be used to study the Taylor series known as the **geometric series** $\sum_{k=0}^{\infty} x^k$ since, according to (5.1), its partial sum $s_n(x)$ is

$$s_n(x) = \sum_{k=0}^{n-1} x^k = \frac{1-x^n}{1-x}. \qquad (x \neq 1) \tag{5.2}$$

Show, from (5.2), that the sequence $s_n(x)$ converges, as $n \to \infty$, for $|x| < 1$, and diverges for $|x| > 1$.
(c) Determine, by any means, the convergence or divergence of the geometric series for the points at the ends of the interval of convergence, $x = \pm 1$. NOTE: The formula (5.2) is quite striking because it reduces $s_n(x)$ to the *closed form* $(1-x^n)/(1-x)$, direct examination of which gives not only the interval of convergence but also the sum function $1/(1-x)$. It is rare that one can reduce $s_n(x)$ to closed form.

6. (a) Derive the Taylor series of $1/(x-1)$ about $x = 4$ using the Taylor series formula (16), and show that your result agrees with (38).
(b) Show that the same result is obtained (more readily) by writing

$$\frac{1}{x-1} = \frac{1}{3+(x-4)} = \frac{1}{3}\frac{1}{1+\frac{x-4}{3}} \tag{6.1}$$

and using the geometric series formula

$$\boxed{\frac{1}{1-t} = \sum_{0}^{\infty} t^n \qquad (|t| < 1)} \tag{6.2}$$

from Exercise 5, with $t = -(x-4)/3$. Further, deduce the x interval of convergence of the result from the convergence condition $|t| < 1$ in (6.2).

7. For each of the following differential equations do the following: Identify $p(x)$ and $q(x)$ and, from them, determine the least possible guaranteed radius of convergence of power series solutions about the specified point x_0; seeking a power series solution form, about that point, obtain the recursion formula and the first four nonvanishing terms in the power series for $y_1(x)$ and $y_2(x)$; verify that y_1, y_2 are LI.

(a) $y'' + 2y' + y = 0, \quad x_0 = 0$
(b) $y'' + 2y' = 0, \quad x_0 = 0$
(c) $y'' + 2y' = 0, \quad x_0 = 3$
(d) $xy'' + y' + y = 0, \quad x_0 = -5$
(e) $xy'' - 2y' + xy = 0, \quad x_0 = 1$
(f) $x^2y'' - y = 0, \quad x_0 = 2$
(g) $xy'' + (3+x)y' + xy = 0, \quad x_0 = -3$
(h) $y'' + y' + (1 + x + x^2)y = 0, \quad x_0 = 0$
(i) $y'' - (1+x^2)y = 0, \quad x_0 = 0$
(j) $y'' - x^3y = 0, \quad x_0 = 0$
(k) $y'' + x^3y' + y = 0, \quad x_0 = 0$
(l) $y'' + xy' + x^2y = 0, \quad x_0 = 0$
(m) $y'' + (x-1)^2y = 0, \quad x_0 = 2$

8. (a)–(m) Use computer software to obtain the general solution, in power series form, for the corresponding problem given in Exercise 7, about the given expansion point.

9. (*Airy equation*) For the **Airy equation**,

$$y'' - xy = 0, \qquad (-\infty < x < \infty) \tag{9.1}$$

derive the power series solution

$$\begin{aligned} y(x) &= a_0y_1(x) + a_1y_2(x) \\ &= a_0\left(1 + \sum_{1}^{\infty} \frac{x^{3n}}{2\cdot 3\cdots(3n-1)(3n)}\right) \\ &\quad + a_1\left(x + \sum_{1}^{\infty} \frac{x^{3n+1}}{3\cdot 4\cdots(3n)(3n+1)}\right) \end{aligned} \tag{9.2}$$

and verify that it is a general solution. NOTE: These series are not summable in closed form in terms of elementary functions thus, certain linear combinations of y_1 and y_2 are adopted as a usable pair of LI solutions. In particular, it turns out to be convenient (for reasons that are not obvious) to use the **Airy functions** $Ai(x)$ and $Bi(x)$, which satisfy these initial conditions: $Ai(0) = 0.35502$, $Ai'(0) = -0.25881$ and $Bi(0) = 0.61493$, $Bi'(0) = 0.44829$.

10. Use computer software to obtain power series solutions of the following initial-value problems, each defined on $0 \leq x < \infty$, through terms of eighth order, and obtain a computer plot of $s_2(x)$, $s_4(x)$, $s_6(x)$, and $s_8(x)$.

(a) $y'' + 4y' + y = 0, \quad y(0) = 1, \quad y'(0) = 0$
(b) $y'' + x^2y = 0, \quad y(0) = 2, \quad y'(0) = 0$
(c) $y'' - xy' + y = 0, \quad y(0) = 0, \quad y'(0) = 1$
(d) $(1+x)y'' + y = 0, \quad y(0) = 2, \quad y'(0) = 0$
(e) $(3+x)y'' + y' + y = 0, \quad y(0) = 0, \quad y'(0) = 1$
(f) $(1+x^2)y'' + y = 0, \quad y(0) = 1, \quad y'(0) = 1$

11. From the given recursion formula alone, determine the radius of convergence of the corresponding power series solutions.

(a) $(n+3)(n+2)a_{n+2} - (n+1)^2a_{n+1} + na_n = 0$
(b) $(n+1)a_{n+2} + 5na_{n+1} + a_n - a_{n-1} = 0$
(c) $(n+1)^2a_{n+2} + (2n^2+1)a_{n+1} - 4a_n = 0$
(d) $(n+1)a_{n+2} - 3(n+2)a_n = 0$
(e) $na_{n+2} + 4na_{n+1} + 3a_n = 0$
(f) $n^2a_{n+2} - 3(n+2)^2a_{n+1} + 3a_{n-1} = 0$

12. In the Comment at the end of Example 6 we wondered what the divergence of the series solution over $7 < x < \infty$

implied about the nature of the solution over that part of the domain. To gain insight, we propose studying a simple problem with similar features. Specifically, consider the problem

$$(x-1)y' + y = 0, \qquad y(4) = 5 \tag{12.1}$$

on the interval $4 \le x < \infty$.
(a) Solve (12.1) analytically, and show that the solution is

$$y(x) = \frac{15}{x-1} \tag{12.2}$$

over $4 \le x < \infty$. Sketch the graph of (12.2), showing it as a solid curve over the domain $4 \le x < \infty$, and dotted over $-\infty < x < 4$.
(b) Solve (12.1), instead, by seeking $y(x) = \sum_0^\infty a_n(x-4)^n$.
(c) Show that the solution obtained in (b) is, in fact, the Taylor expansion of (12.2) about $x = 4$ and that it converges only in $|x-4| < 3$ so that it represents the solution (12.2) only over the $4 \le x < 7$ part of the domain, even though the solution (12.2) exists and is perfectly well-behaved over $7 < x < \infty$.

13. Rework Example 5 without using the $\sum$ summation notation. That is, just write out the series, as we did in the introductory example of Section 4.1. Keep powers of x up to and including fifth order, x^5, and show that your result agrees (up to terms of fifth order) with that given in (29).

14. Rework Example 6 without using the $\sum$ summation notation. That is, just write out the series as we did in the introductory example of Section 4.1. Keep powers of $x - 4$ up to and including fourth order $(x-4)^4$, and show that your result agrees (up to terms of fourth order) with that given in (46).

15. (*Cesàro summability*) Although (5) gives the usual definition of infinite series, it is not the only possible one nor the only one used. For example, according to **Cesàro summability**, which is especially useful in the theory of Fourier series, one defines

$$\sum_1^\infty a_n \equiv \lim_{N\to\infty} \frac{s_1 + s_2 + \cdots + s_N}{N}, \tag{15.1}$$

that is, the limit of the arithmetic means of the partial sums. It can be shown that if a series converges to s according to "ordinary convergence" [equation (5)], then it will also converge to the same value in the Cesàro sense. Yet, there are series that diverge in the ordinary sense but that converge in the Cesàro sense. Show that for the geometric series (see Exercise 5),

$$\frac{s_1 + s_2 + \cdots + s_N}{N} = \frac{1}{1-x} - \frac{x}{N}\frac{1-x^N}{(1-x)^2} \tag{15.2}$$

for all $x \ne 1$, and use that result to show that the Cesáro definition gives divergence for all $|x| > 1$ and for $x = 1$, and convergence for $|x| < 1$, as does ordinary convergence, but that for $x = -1$ it gives convergence to $1/2$, whereas according to ordinary convergence the series diverges for $x = -1$.

4.3 The Method of Frobenius

4.3.1. Singular points. In this section we continue to consider series solutions of the equation

$$y'' + p(x)y' + q(x)y = 0. \tag{1}$$

From Section 4.2, we know that we can find two LI solutions as power series expansions about any point x_0 at which both p and q are analytic. We call such a point x_0 an **ordinary point** of the equation (1). Typically, p and q are analytic everywhere on the x axis except perhaps at one or more singular points, so that all points of the x axis, except perhaps a few, are ordinary points. In that case one can readily select such an x_0 and develop two LI power series solutions about that point.

Nevertheless, in the present section we examine singular points more closely, and show that one can often obtain modified series solutions about singular points.

Why should we want to develop a method of finding series solutions about a singular point when we can stay away from the singular point and expand about an ordinary point? There are at least two reasons, which are explained later in this section.

Proceeding, we begin by classifying singular points as follows:

DEFINITION 4.3.1 *Regular and Irregular Singular Points of (1)*
Let x_0 be a singular point of p and/or q. We classify it as a regular or irregular singular point of equation (1) as follows: x_0 is
(a) a **regular singular point of** (1) if $(x - x_0)p(x)$ and $(x - x_0)^2 q(x)$ are analytic at x_0,
(b) an **irregular singular point** of (1) if it is not a regular singular point.

EXAMPLE 1. Consider $x(x-1)^2 y'' - 3y' + 5y = 0$ or, dividing by $x(x-1)^2$ to put the equation in the standard form $y'' + p(x)y' + q(x)y = 0$,

$$y'' - \frac{3}{x(x-1)^2} y' + \frac{5}{x(x-1)^2} y = 0. \tag{2}$$

Thus, $p(x) = -3/[x(x-1)^2]$ and $q(x) = 5/[x(x-1)^2]$. These are analytic for all x except for $x = 0$ and $x = 1$, so every x is an ordinary point except for those points. Let us classify those two singular points:

$$\begin{aligned}
x_0 = 0: \quad & (x - x_0)p(x) = (x - 0)\left(-\frac{3}{x(x-1)^2}\right) = -\frac{3}{(x-1)^2} \\
& (x - x_0)^2 q(x) = (x - 0)^2 \left(\frac{5}{x(x-1)^2}\right) = \frac{5x}{(x-1)^2} \\
x_0 = 1: \quad & (x - x_0)p(x) = (x - 1)\left(-\frac{3}{x(x-1)^2}\right) = -\frac{3}{x(x-1)} \\
& (x - x_0)^2 q(x) = (x - 1)^2 \left(\frac{5}{x(x-1)^2}\right) = \frac{5}{x}.
\end{aligned} \tag{3a,b,c,d}$$

To classify the singular point at $x = 0$, consider (3a) and (3b). Since the right-hand sides of (3a) and (3b) are analytic* at 0, we classify $x = 0$ as a regular singular point of (2). (The fact that those right-hand sides are singular elsewhere, at $x = 1$, is irrelevant.) To classify the singular point at $x = 1$, we turn to (3c) and (3d). Whereas the right-hand side of (3d) is analytic at $x = 1$, the right-hand side of (3c) is not, so we classify the singular point at $x = 1$ as an irregular singular point of (2). ∎

EXAMPLE 2. Consider the case

$$y'' + \sqrt{x}\, y = 0. \qquad (0 \le x < \infty) \tag{4}$$

*Recall the rule of thumb given in the last sentence of Section 4.2.1, that we will classify a function as analytic at a given point if it is infinitely differentiable at that point.

Then $p(x) = 0$ and $q(x) = \sqrt{x}$, and these are analytic (infinitely differentiable) for all $x > 0$, but not at $x = 0$ because $q(x)$ is not even once differentiable there, let alone infinitely differentiable. To classify the singular point at $x = 0$, observe that $(x - x_0)p(x) = (x)(0) = 0$ is analytic at $x = 0$, but $(x - x_0)^2 q(x) = x^2\sqrt{x} = x^{5/2}$ is not; it is twice differentiable there (those derivatives being zero), but all higher derivatives are undefined at $x = 0$. Thus, $x = 0$ is an irregular singular point of (4). (See Exercise 2.) ∎

4.3.2. Method of Frobenius. To develop the method of Frobenius, we require that the singular point about which we expand be a *regular* singular point. Before stating theorems and working examples, let us motivate the idea behind the method. We consider the equation

$$y'' + p(x)y' + q(x)y = 0 \tag{5}$$

to have a regular singular point at the origin (and perhaps other singular points as well). There is no loss of generality in assuming it to be at the origin since, if it is at $x = x_0 \neq 0$, we can always make a change of variable $\xi = x - x_0$ to move it to the origin in terms of the new variable ξ (Exercise 3). Until stated otherwise, let us assume that the interval of interest is $\boldsymbol{x > 0}$.

We begin by multiplying equation (5) by x^2 and rearranging terms as

$$x^2y'' + x\left[xp(x)\right]y' + \left[x^2q(x)\right]y = 0. \tag{6}$$

Since $x = 0$ is a regular singular point, it follows that xp and x^2q can be expanded about the origin in convergent Taylor series, so we can write

$$x^2y'' + x\left(p_0 + p_1x + \cdots\right)y' + \left(q_0 + q_1x + \cdots\right)y = 0. \tag{7}$$

Locally, in the neighborhood of $x = 0$, we can approximate (7) as

$$x^2y'' + p_0xy' + q_0y = 0, \tag{8}$$

which is a Cauchy-Euler equation. As such, (8) has at least one solution in the form x^r, for some constant r. Returning to (7), it is reasonable to expect that equation, likewise, to have at least one solution that behaves like x^r (for the same value of r) in the neighborhood of $x = 0$. More completely, we expect it to have at least one solution of the form

$$y(x) = x^r\left(a_0 + a_1x + a_2x^2 + \cdots\right), \tag{9}$$

where the power series factor is needed to account for the deviation of $y(x)$, away from $x = 0$, from its asymptotic behavior $y(x) \sim a_0x^r$ as $x \to 0$. That is, in place of the power series expansion

$$y(x) = \sum_{0}^{\infty} a_nx^n \tag{10}$$

that is guaranteed to work when $x = 0$ is an ordinary point of (5), it appears that we should seek $y(x)$ in the more general form

$$\boxed{y(x) = x^r \sum_0^\infty a_n x^n = \sum_0^\infty a_n x^{n+r}} \tag{11}$$

if $x = 0$ is a regular singular point. Is (11) really different from (10)? Yes, because whereas (10) necessarily represents an analytic function, (11) represents a nonanalytic function because of the x^r factor (unless r is a nonnegative integer).

Let us try the solution form (11) in an example.

EXAMPLE 3. The equation

$$6x^2y'' + 7xy' - (1 + x^2)y = 0, \qquad (0 < x < \infty) \tag{12}$$

has a regular singular point at $x = 0$ because whereas $p(x) = 7x/(6x^2) = 7/(6x)$ and $q(x) = -(1 + x^2)/(6x^2)$ are singular at $x = 0$, $xp(x) = 7/6$ and $x^2q(x) = -(1 + x^2)/6$ are analytic there. Let us seek $y(x)$ in the form (11). Putting that form into (12) gives

$$6x^2 \sum_0^\infty (n+r)(n+r-1)a_n x^{n+r-2} + 7x \sum_0^\infty (n+r)a_n x^{n+r-1} \\ -(1+x^2)\sum_0^\infty a_n x^{n+r} = 0 \tag{13}$$

or

$$\sum_0^\infty 6(n+r)(n+r-1)a_n x^{n+r} + \sum_0^\infty 7(n+r)a_n x^{n+r} \\ - \sum_0^\infty a_n x^{n+r} - \sum_0^\infty a_n x^{n+r+2} = 0. \tag{14}$$

Letting $n + r + 2 = m + r$ in the last sum, (14) becomes

$$\sum_0^\infty [6(n+r)(n+r-1) + 7(n+r) - 1]\, a_n x^{n+r} - \sum_2^\infty a_{m-2} x^{m+r} = 0. \tag{15}$$

Changing the lower limit in the last sum to 0, with the understanding that $a_{-2} = a_{-1} \equiv 0$, and changing the m's to n's, we can finally combine the sums as

$$\sum_0^\infty \left\{ \left[6(n+r)^2 + n + r - 1\right] a_n - a_{n-2} \right\} x^{n+r} = 0, \tag{16}$$

where we have also simplified the square bracket in (15) to $6(n + r)^2 + n + r - 1$. From (16), we infer the recursion formula

$$\left[6(n+r)^2 + n + r - 1\right] a_n - a_{n-2} = 0 \tag{17}$$

for each $n = 0, 1, 2, \ldots$.

In particular, $n = 0$ gives

$$\left(6r^2 + r - 1\right) a_0 - a_{-2} = 0 \tag{18}$$

or, since $a_{-2} \equiv 0$,

$$\left(6r^2 + r - 1\right) a_0 = 0. \tag{19}$$

Observe carefully that we may assume, with no loss of generality, that $a_0 \neq 0$ for, if $a_0 = 0$ in (9) and $a_1 \neq 0$, then we can factor an x out of the series and absorb it into the x^r. Rather, let us assume, in writing (9) that all such factors have already been absorbed into the x^r, and that a_0 is the first nonvanishing coefficient.

Proceeding, with $a_0 \neq 0$, it follows from (19) that

$$6r^2 + r - 1 = 0 \tag{20}$$

so $r = -1/2$ and $1/3$. We expect each of the choices, $r = -1/2$ and $r = 1/3$, to lead to a Frobenius-type solution (11), and the two solutions thus obtained to be LI. Let us see. First, set $r = -1/2$. The corresponding recursion formula (17) is then

$$a_n = \frac{1}{6\left(n - \frac{1}{2}\right)^2 + n - \frac{3}{2}} a_{n-2} \tag{21}$$

for $n = 1, 2, \ldots$, since the $n = 0$ case has already been carried out:

$$\begin{aligned}
n = 1: &\quad a_1 = a_{-1} = 0, \\
n = 2: &\quad a_2 = \frac{1}{14} a_0, \\
n = 3: &\quad a_3 = \frac{1}{39} a_1 = 0, \\
n = 4: &\quad a_4 = \frac{1}{76} a_2 = \frac{1}{(76)(14)} a_0, \\
n = 5: &\quad a_5 = \frac{1}{125} a_3 = 0, \\
n = 6: &\quad a_6 = \frac{1}{186} a_4 = \frac{1}{(186)(76)(14)} a_0,
\end{aligned} \tag{22}$$

and so on. From these results we have the solution

$$\begin{aligned}
y(x) &= a_0 x^{-1/2} \left[1 + \frac{1}{14} x^2 + \frac{1}{(76)(14)} x^4 + \frac{1}{(186)(76)(14)} x^6 + \cdots\right] \\
&= a_0 y_1(x),
\end{aligned} \tag{23}$$

where a_0 remains arbitrary.

Next, set $r = 1/3$. The corresponding recursion formula (17) is then

$$a_n = \frac{1}{6\left(n + \frac{1}{3}\right)^2 + n - \frac{2}{3}} a_{n-2}, \tag{24}$$

and proceeding as we did for $r = -1/2$, we obtain the solution (Exercise 4)

$$\begin{aligned} y(x) &= a_0 x^{1/3}\left[1 + \frac{1}{34}x^2 + \frac{1}{(116)(34)}x^4 + \frac{1}{(246)(116)(34)}x^6 + \cdots\right] \\ &= a_0 y_2(x), \end{aligned} \tag{25}$$

where a_0 remains arbitrary. [Of course, the a_0's in (23) and (25) have nothing to do with each other; they are independent arbitrary constants.] According to Theorem 3.2.4, the solutions y_1 and y_2 are LI because neither is a scalar multiple of the other, so a general solution to (12) is $y(x) = C_1 y_1(x) + C_2 y_2(x)$, where y_1 and y_2 are given in (23) and (25).

What are the regions of convergence of the series in (23) and (25)? Though we don't have the general terms for those series, we can use their recursion formulas, (21) and (24), respectively, to study their convergence. Consider the series in (23) first. Its recursion formula is (21) or, equivalently,

$$a_{n+2} = \frac{1}{6\left(n + 2 - \frac{1}{2}\right)^2 + n + 2 - \frac{3}{2}} a_n. \tag{26}$$

We need to realize that the a_{n+2} on the left side is really the next coefficient after a_n, the "a_{n+1}" in Theorem 4.2.2, since every other term in the series is missing (because $a_1 = a_3 = a_5 = \cdots = 0$). Thus, (26) gives

$$\lim_{n\to\infty}\left|\frac{\text{“}a_{n+1}\text{”}}{a_n}\right| = \lim_{n\to\infty}\frac{1}{6\left(n + \frac{3}{2}\right)^2 + n + \frac{1}{2}} = 0, \tag{27}$$

and it follows from Theorem 4.2.2 that $R = \infty$; the series converges for all x. Of course, the $x^{-1/2}$ factor in (23) "blows up" at $x = 0$, so (23) is valid in $0 < x < \infty$, which is the full interval specified in (12).

Similarly, we can show that the series in (24) converges for all x, so (25) is valid over the full interval $0 < x < \infty$. ■

With Example 3 completed, we can come back to the important question that we posed near the beginning of Section 4.3.1: "Why should we want to develop a method of finding series solutions about a singular point when we can stay away from the singular point and expand about an ordinary point?". Observe that our Frobenius-type solution $y(x) = C_1 y_1(x) + C_2 y_2(x)$, with $y_1(x)$ and $y_2(x)$ given by (23) and (25), was valid on the full interval $0 < x < \infty$. Furthermore, it even showed us the singular behavior at the origin explicitly:

$$\begin{aligned} y(x) &= C_1 x^{-1/2}\left(1 + \frac{1}{14}x^2 + \cdots\right) + C_2 x^{1/3}\left(1 + \frac{1}{34}x^2 + \cdots\right) \\ &\sim C_1 \frac{1}{\sqrt{x}} \end{aligned} \tag{28}$$

as $x \to 0$. In contrast, if we had avoided the singular point $x = 0$ and pursued power series expansions about an ordinary point, say $x = 2$, then the resulting solution would have been valid only in $0 < x < 4$, and it would not have explicitly displayed the $1/\sqrt{x}$ singular behavior at $x = 0$.

Let us review the ideas presented above, and get ready to state the main theorem. If $x = 0$ is a regular singular point of the equation $y'' + p(x)y' + q(x) = 0$, which we rewrite as

$$x^2y'' + x\left[xp(x)\right]y' + \left[x^2q(x)\right]y = 0, \tag{29}$$

then $xp(x)$ and $x^2q(x)$ admit Taylor series representations $xp(x) = p_0 + p_1x + \cdots$ and $x^2q(x) = q_0 + q_1x + \cdots$. Locally then, near $x = 0$, (29) can be approximated as

$$x^2y'' + p_0xy' + q_0y = 0, \tag{30}$$

which is a Cauchy-Euler equation. Cauchy-Euler equations, we recall, always admit at least one solution in the form x^r, and this fact led us to seek solutions $y(x)$ to (29) that behave like $y(x) \sim x^r$ as $x \to 0$, or

$$y(x) = x^r\sum_0^\infty a_nx^n = \sum_0^\infty a_nx^{n+r}, \tag{31}$$

where the $\sum_0^\infty a_nx^n$ factor is to account for the deviation of y from the local behavior $y(x) \sim x^r$ away from $x = 0$. Putting (31) into

$$x^2y'' + x\left(p_0 + p_1x + \cdots\right)y' + \left(q_0 + q_1x + \cdots\right)y = 0 \tag{32}$$

gives

$$\begin{aligned}\sum_0^\infty(n+r)(n+r-1)a_nx^{n+r} + (p_0 + p_1x + \cdots)\sum_0^\infty(n+r)a_nx^{n+r}&\\ + (q_0 + q_1x + \cdots)\sum_0^\infty a_nx^{n+r} &= 0,\end{aligned} \tag{33}$$

and equating coefficients of the various powers of x to zero gives

$$\begin{aligned}
x^r &: \left[r(r-1) + p_0r + q_0\right]a_0 = 0, && \text{(34a)}\\
x^{r+1} &: \left[(r+1)r + p_0(r+1) + q_0\right]a_1 + (p_1r + q_1)a_0 = 0, && \text{(34b)}\\
x^{r+2} &: \left[(r+2)(r+1) + p_0(r+2) + q_0\right]a_2 + (\text{etc})a_1\\
&\quad +(\text{etc})a_0 = 0, && \text{(34c)}\\
x^{r+3} &: \left[(r+3)(r+2) + p_0(r+3) + q_0\right]a_3 + (\text{etc})a_2 + (\text{etc})a_1\\
&\quad +(\text{etc})a_0 = 0, && \text{(34d)}
\end{aligned}$$

and so on, where we've used "etc's" for brevity since we're most interested, here, in showing the *form* of the equations. Assuming, without loss of generality, that $a_0 \neq 0$, (34a) gives

$$\boxed{r^2 + (p_0 - 1)r + q_0 = 0,} \tag{35}$$

which quadratic equation for r is called the **indicial equation**; in Example 3 the indicial equation was equation (20). Let the roots be r_1 and r_2. Setting $r = r_1$ in (34b,c,d,...) gives a system of linear algebraic equations to find $a_1, a_2, \ldots$ in terms of a_0, with a_0 remaining arbitrary. Next, we set $r = r_2$ in (34b,c,d,...) and again try to solve for $a_1, a_2, \ldots$ in terms of a_0. If all goes well, those steps should produce two LI solutions of the differential equation $y'' + p(x)y' + q(x) = 0$.

The process is straightforward and was carried out successfully in Example 3. Can anything go wrong? Yes. One potential difficulty is that *the indicial equation might have repeated roots* ($r_1 = r_2$), in which case the procedure gives only one solution. To seek guidance as to how to find a second LI solution, realize that the same situation occurred for the simplified problem, the Cauchy-Euler equation (30): if, seeking $y(x) = x^r$ in (30), we obtain a repeated root for r, then a second solution can be found (by the method of reduction of order) to be of the form x^r times $\ln x$. Similarly, for $y'' + p(x)y' + q(x) = 0$, as we shall see in the theorem below [specifically, (41b)].

The other possible difficulty, which is more subtle, occurs *if the roots differ by a nonzero integer*. For observe that if we denote the bracketed coefficient of a_0 in (34a) as $F(r)$, then the coefficient of a_1 in (34b) is $F(r+1)$, that of a_2 in (34c) is $F(r+2)$, and so on. To illustrate, suppose that $r_1 = r_2 + 1$, so that the roots differ by 1. Then not only will $F(r)$ vanish in (34a) when we are using $r = r_2$, but so will $F(r+1)$ in (34b) [though not $F(r+2)$ in (34c), nor $F(r+3)$ in (34d), etc.], in which case (34b) becomes $0a_1 + (p_1 r_2 + q_1)a_0 = 0$. If $p_1 r_2 + q_1$ happens not to be zero then the equation (34b) cannot be satisfied, and there is no set of a_n's that satisfy the system (34). Thus, for the algebraically smaller root r_2 (e.g., -6 is algebraically smaller than 2), no solution is found. But if $p_1 r_2 + q_1$ does equal zero, then (34b) becomes $0a_1 = 0$ and a_1 (in addition to a_0) remains arbitrary. Then (34c,d,...) give $a_2, a_3, \ldots$ as linear combinations of a_0 and a_1, and one obtains a general solution

$$\begin{aligned} y(x) &= a_0 x^{r_2} \text{ (a power series) } + a_1 x^{r_2} \text{ (a different power series)} \\ &= a_0 y_1(x) + a_1 y_2(x), \end{aligned} \tag{36}$$

where a_0, a_1 are arbitrary and y_1, y_2 are LI.

If, however, r_2 gives no solution, then we can turn to r_1. For r_1 the difficulty cited in the preceding paragraph does not occur, and the method produces a single solution "$y_2(x)$."

If, instead, $r_1 = r_2 + 2$, say, then the same sort of difficulty shows up, but not until (34c). Similarly, if $r_1 = r_2 + 3$, $r_1 = r_2 + 4$, and so on.

The upshot is that if r_1, r_2 differ by a nonzero integer, then the algebraically smaller root r_2 leads either to no solution or to a general solution. In either case, the larger root r_1 leads to one solution.

The theorem is as follows.

THEOREM 4.3.1 *Regular Singular Point; Frobenius Solution*
Let $x = 0$ be a regular singular point of the differential equation

$$y'' + p(x)y' + q(x)y = 0, \qquad (x > 0) \tag{37}$$

with $xp(x) = p_0 + p_1 x + \cdots$ and $x^2 q(x) = q_0 + q_1 x + \cdots$ having radii of convergence R_1, R_2 respectively. Let r_1, r_2 be the roots of the indicial equation

$$r^2 + (p_0 - 1)r + q_0 = 0, \tag{38}$$

where $r_1 \geq r_2$ if the roots are real. (Otherwise they are complex conjugates.) Seeking $y(x)$ in the form

$$y(x) = x^r \sum_0^\infty a_n x^n = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0) \tag{39}$$

with $r = r_1$ inevitably leads to a solution

$$y_1(x) = x^{r_1} \sum_0^\infty a_n x^n, \qquad (a_0 \neq 0) \tag{40}$$

where $a_1, a_2, \dots$ are known multiples of a_0, which remains arbitrary. For definiteness, we choose $a_0 = 1$ in (40). The form of the second LI solution, $y_2(x)$, depends on r_1 and r_2 as follows:

(i) r_1 and r_2 *distinct and not differing by an integer.* (Complex conjugate roots belong to this case.) Then with $r = r_2$, (39) yields

$$y_2(x) = x^{r_2} \sum_0^\infty b_n x^n, \qquad (b_0 \neq 0) \tag{41a}$$

where the b_n's are generated by the same recursion relation as the a_n's, but with $r = r_2$ instead of $r = r_1$; $b_1, b_2, \dots$ are known multiples of b_0, which is arbitrary. For definiteness, we choose $b_0 = 1$ in (41a).
(ii) *Repeated roots*, $r_1 = r_2 \equiv r$. Then $y_2(x)$ can be found in the form

$$y_2(x) = y_1(x) \ln x + x^r \sum_1^\infty c_n x^n. \tag{41b}$$

(iii) $r_1 - r_2$ *equal to an integer.* Then the smaller root r_2 leads to both solutions, $y_1(x)$ and $y_2(x)$, or to neither. In either case, the larger root r_1 gives the single solution (40). In the latter case, $y_2(x)$ can be found in the form

$$y_2(x) = \kappa y_1(x) \ln x + x^{r_2} \sum_0^\infty d_n x^n, \tag{41c}$$

where the constant κ may turn out to be zero, in which case there is no logarithmic term in (41c).

The radius of convergence of each of the series in (40) and (41) is at least as large as the smaller of R_1, R_2.

If (37) is on $x < 0$ rather than $x > 0$, then the foregoing is valid, provided that each x^r, x^{r_1}, x^{r_2} and $\ln x$ is changed to $|x|^r$, $|x|^{r_1}$, $|x|^{r_2}$ and $\ln|x|$, respectively.

Outline of Proof of (ii): Our discussion preceding this theorem contained an outline of the proof of case (i), and also some discussion of case (iii). Here, let us focus on case (ii) and, again, outline the main ideas behind a proof. We consider the case of repeated roots, where $r_1 = r_2 \equiv r$. Since $y_1(x)$, given by (40), is a solution, then so is $y(x) = Ay_1(x)$, where A is arbitrary. To find $y_2(x)$, let us use reduction of order; that is, seek $y_2(x) = A(x)y_1(x)$, where $y_1(x)$ is known and $A(x)$ is to be found. Putting that form into (37) gives

$$A''y_1 + A'\left(2y_1' + py_1\right) + A\left(y_1'' + py_1' + qy_1\right) = 0. \tag{42}$$

Since y_1 satisfies (37), the last term in (42) is zero, so (42) becomes

$$A''y_1 + A'\left(2y_1' + py_1\right) = 0. \tag{43}$$

Replacing A'' by dA'/dx, multiplying through by dx, and dividing by y_1 and A' gives

$$\frac{dA'}{A'} + 2\frac{dy_1}{y_1} + p\,dx = 0. \tag{44}$$

Integrating,

$$\ln|A'| + 2\ln|y_1| + \int p(x)\,dx = \text{constant, say } \ln C, \text{ for } C > 0,$$

so

$$\ln\frac{|A'|\,|y_1|^2}{C} = -\int p(x)\,dx,$$

and

$$\begin{aligned}|A'(x)| &= C\,\frac{e^{-\int p(x)\,dx}}{y_1^2(x)} = C\,\frac{e^{-\int\left(\frac{p_0}{x}+p_1+p_2x+\cdots\right)dx}}{\left[x^r\left(1+a_1x+\cdots\right)\right]^2}\\ &= C\,\frac{e^{-p_0\ln x}e^{\left(-p_1x-\cdots\right)}}{x^{2r}\left(1+2a_1x+\cdots\right)},\end{aligned} \tag{45}$$

where we write $\ln x$ rather than $\ln|x|$ since $x > 0$ here, by assumption.

Since $\exp\left(-\int p(x)\,dx\right) > 0$, we see from the first equality in (45) that $A'(x)$ is either everywhere positive or everywhere negative on the interval. Thus, we can drop the absolute value signs around $A'(x)$ if we now allow C to be positive or negative. Further, $e^{-p_0 \ln x} = e^{\ln x^{-p_0}} = x^{-p_0}$, and $e^{(-p_1 x - \cdots)}/(1 + 2a_1 x + \cdots)$ is analytic at $x = 0$ and can be expressed in Taylor series form as $1 + \kappa_1 x + \kappa_2 x^2 + \cdots$, so

$$A'(x) = C \frac{1}{x^{2r+p_0}} (1 + \kappa_1 x + \cdots). \tag{46}$$

For r to be a double root of the indicial equation (38), it is necessary that $2r + p_0 = 1$, in which case integration of (46) gives

$$A(x) = C(\ln x + \kappa_1 x + \cdots). \tag{47}$$

Finally, setting $C = 1$ with no loss of generality, we have the form

$$\begin{aligned} y_2(x) &= A(x) y_1(x) = (\ln x + \kappa_1 x + \cdots) y_1(x) \\ &= y_1(x) \ln x + (\kappa_1 x + \cdots) x^r (1 + a_1 x + \cdots) \\ &= y_1(x) \ln x + x^r \sum_1^\infty c_n x^n, \end{aligned} \tag{48}$$

as given in (41b). ■

In short, the Frobenius method is guaranteed to provide two LI solutions to (37) if $x = 0$ is a regular singular point of that equation. If $x = 0$ is an irregular singular point, the theorem does not apply. That case is harder, and we have no analogous theory to guide us.

EXAMPLE 4. *Case (ii).* Solve the equation

$$x^2 y'' - (x + x^2) y' + y = 0; \qquad (0 < x < \infty) \tag{49}$$

that is, find two LI solutions. The only singular point of (49) is $x = 0$, and it is a regular singular point. Seeking

$$y(x) = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0) \tag{50}$$

substitution of that form into (49) gives

$$\begin{aligned} &\sum_0^\infty (n+r)(n+r-1) a_n x^{n+r} - \sum_0^\infty (n+r) a_n x^{n+r} \\ &\qquad - \sum_0^\infty (n+r) a_n x^{n+r+1} + \sum_0^\infty a_n x^{n+r} = 0. \end{aligned} \tag{51}$$

Set $n + 1 = m$ in the third sum, change the lower limit from $n = 0$ to $m = 1$, extend that limit back to 0 by defining $a_{-1} \equiv 0$, change the m's back to n's, and combine the four sums. Those steps give

$$\sum_0^\infty \{[(n+r)(n+r-1) - (n+r) + 1] a_n - (n+r-1) a_{n-1}\} x^{n+r} = 0$$

and hence the recursion formula

$$[(n+r)(n+r-1)-(n+r)+1]\,a_n-(n+r-1)a_{n-1}=0, \tag{52}$$

for $n=0,1,2,\ldots$. For $n=0$, (52) becomes $(r^2-2r+1)a_0-(r-1)a_{-1}=0$. Since $a_{-1}\equiv 0$ and $a_0\neq 0$, the latter gives the indicial equation

$$r^2-2r+1=0, \tag{53}$$

with repeated roots $r=1,1$. Thus, this example illustrates case (ii). Putting $r=1$ into (52), we obtain the recursion formula

$$a_n=\frac{1}{n}a_{n-1} \tag{54}$$

for $n=1,2,\ldots$. Thus, $a_1=a_0$, $a_2=\frac{1}{2}a_1=\frac{1}{2}a_0$, $a_3=\frac{1}{3}a_2=\frac{1}{3!}a_0$ and we can see that $a_n=\frac{1}{n!}a_0$, so

$$\begin{aligned} y(x)&=x\left(a_0+\frac{a_0}{1!}x+\frac{a_0}{2!}x^2+\cdots\right)\\ &=a_0\sum_0^\infty\frac{x^{n+1}}{n!}=a_0y_1(x). \end{aligned} \tag{55}$$

In this case we can identify the series as xe^x, but we are not always this fortunate, so let us keep working with the series form in this example.

Theorem 4.3.1 tells us that y_2 can be found in the form (41b), where $r=1$ and the c_n's are to be determined. Putting that form into (49) gives

$$\begin{aligned} x^2y_2''-(x+x^2)y_2'+y_2&=\left[x^2y_1''-(x+x^2)y_1'+y_1\right]\ln x+2xy_1'-(2+x)y_1\\ &\quad+\sum_1^\infty n(n+1)c_nx^{n+1}+\sum_1^\infty c_nx^{n+1}\\ &\quad-\sum_1^\infty(n+1)c_nx^{n+1}-\sum_1^\infty(n+1)c_nx^{n+2}=0, \end{aligned} \tag{56}$$

where $y_1(x)=\sum_0^\infty x^{n+1}/n!$. The square-bracketed terms in (56) cancel to zero because y_1 is a solution of (49). If we move $2xy_1'-(2+x)y_1$ to the right-hand side, and write out the various series, then (56) becomes

$$c_1x^2+4c_2x^3+9c_3x^4+\cdots-2c_1x^3-3c_2x^4-\cdots=-x^2-x^3-\frac{1}{2}x^4-\cdots, \tag{57}$$

and equating coefficients of like powers of x, on the left- and right-hand sides, gives

$$\begin{aligned} x^2:&\quad c_1=-1,\\ x^3:&\quad 4c_2-2c_1=-1,\\ x^4:&\quad 9c_3-3c_2=-\tfrac{1}{2}, \end{aligned} \tag{58}$$

and so on. Thus, $c_1 = -1$, $c_2 = -\frac{3}{4}$, $c_3 = -\frac{11}{36}, \ldots$ and

$$\begin{aligned} y_2(x) &= y_1(x) \ln x + \sum_1^\infty c_n x^{n+1} \\ &= \left(x + x^2 + \frac{1}{2}x^3 + \frac{1}{6}x^4 + \cdots \right) \ln x - x^2 - \frac{3}{4}x^3 - \frac{11}{36}x^4 - \cdots . \end{aligned} \tag{59}$$

If, instead, we retain the summation notation, then in place of (57) we obtain, after manipulation and simplification,

$$\sum_1^\infty \left(n^2 c_n - n c_{n-1} \right) x^{n+1} = - \sum_1^\infty \frac{1}{(n-1)!} x^{n+1} \tag{60}$$

[where $c_0 \equiv 0$ because there is no c_0 term in (41b)] and hence the recursion formula

$$n^2 c_n - n c_{n-1} = - \frac{1}{(n-1)!}$$

or, more conveniently,

$$c_n = \frac{1}{n} c_{n-1} - \frac{1}{n\, n!}. \qquad (c_0 \equiv 0) \tag{61}$$

Solving (61) gives

$$\begin{aligned} c_1 &= -1, \\ c_2 &= -\frac{1}{2!} \left(1 + \frac{1}{2} \right) = -\frac{3}{4}, \\ c_3 &= -\frac{1}{3!} \left(1 + \frac{1}{3} \right) = -\frac{11}{36}, \end{aligned} \tag{62}$$

and so on. These results agree with those obtained from (58), but the gain, here, is that (61) can give us as many c_n's as we wish. In fact, by carrying (62) further we can see that

$$c_n = -\frac{1}{n!} \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} \right) \tag{63}$$

for any $n = 1, 2, \ldots$. [The price that we paid for that gain was that we needed to manipulate the series by shifting some of the summation indices and summation limits in order to obtain (60).]

COMMENT 1. In this example we were able to sum the series in (55) and obtain $y_1(x)$ in the closed form

$$y_1(x) = x e^x. \tag{64}$$

In such a case it is more convenient to seek y_2 by actually carrying out the reduction-of-order process that lay behind the general form (41b). Thus, seek $y_2(x) = A(x) y_1(x)$. The steps were already carried out in the preceding outline of the proof of Theorem 4.3.1, so we can immediately use (45). With $C = 1$,

$$A'(x) = \frac{e^{-\int p(x)\, dx}}{y_1^2(x)} - \frac{e^{-\int (-\frac{1}{x} - 1) dx}}{x^2 e^{2x}} = \frac{e^{\ln x} e^x}{x^2 e^{2x}} = \frac{e^{-x}}{x}, \tag{65}$$

so

$$\begin{aligned} A(x) &= \int \frac{e^{-x}}{x}\,dx = \int \left(\frac{1}{x} - 1 + \frac{x}{2!} - \frac{x^2}{3!} + \cdots\right) dx \\ &= \ln x + \int \sum_1^\infty (-1)^n \frac{x^{n-1}}{n!}\,dx = \ln x + \sum_1^\infty (-1)^n \frac{x^n}{n\,n!} \end{aligned} \tag{66}$$

and

$$y_2(x) = A(x)y_1(x) = \left[\ln x + \sum_1^\infty (-1)^n \frac{x^n}{n\,n!}\right] xe^x, \tag{67}$$

which expression is found, upon expanding the e^x, to agree with (59).

COMMENT 2. As a matter of fact, we can leave the integral in (66) intact because it is a tabulated function. Specifically, the integral of e^{-x}/x, though nonelementary, comes up often enough for it to have been defined, in the literature, as a special function:

$$\boxed{E_1(x) = \int_x^\infty \frac{e^{-t}}{t}\,dt} \qquad (x > 0) \tag{68}$$

is known as the **exponential integral**. Among its various properties is the expansion

$$E_1(x) = -\gamma - \ln x - \sum_1^\infty (-1)^n \frac{x^n}{n\,n!}, \qquad (x > 0) \tag{69}$$

where $\gamma = 0.5772157$ is **Euler's constant**. Using the $E_1(x)$ function, we can express

$$\begin{aligned} y_2(x) &= A(x)y_1(x) = \left(\int_a^x \frac{e^{-t}}{t}\,dt\right) xe^x \\ &= \left(\int_a^\infty \frac{e^{-t}}{t}\,dt - \int_x^\infty \frac{e^{-t}}{t}\,dt\right) xe^x = [E_1(a) - E_1(x)]\,xe^x, \end{aligned} \tag{70}$$

for any $a > 0$. The $E_1(a)xe^x$ term is merely $E_1(a)$ times $y_1(x)$, so it can be dropped with no loss. Further, the factor -1 in front of the $E_1(x)xe^x$ can likewise be dropped. Thus, in this example we were able to obtain both solutions in closed form, $y_1(x) = xe^x$ and $y_2(x) = E_1(x)xe^x$.

COMMENT 3. Observe that the Taylor series

$$xp(x) = x\left(\frac{-x - x^2}{x^2}\right) = -1 - x, \qquad x^2q(x) = x^2\left(\frac{1}{x^2}\right) = 1 \tag{71}$$

both terminate, hence they surely converge with $R_1 = \infty$ and $R_2 = \infty$, respectively. Thus, Theorem 4.3.1 guarantees that $\sum_0^\infty a_n x^n$ in (40) and $\sum_0^\infty c_n x^n$ in (41b) will likewise have infinite radii of convergence. Of course the $\ln x$ in (41b) tends to $-\infty$ as $x \to 0$, but nevertheless our solutions y_1 and y_2 are indeed valid on the full interval $0 < x < \infty$. ▮

EXAMPLE 5. *Case (iii).* Solve the equation

$$xy'' + y = 0; \qquad (0 < x < \infty) \tag{72}$$

that is, find two LI solutions. The only singular point of (72) is $x = 0$, and it is a regular singular point. Seeking

$$y(x) = \sum_0^\infty a_n x^{n+r}, \qquad (a_0 \neq 0)$$

(72) becomes

$$\sum_0^\infty (n+r)(n+r-1)a_n x^{n+r-1} + \sum_0^\infty a_n x^{n+r} = 0. \tag{73}$$

Set $n - 1 = m$ in the first sum, in which case the lower summation limit becomes -1, then change the m's back to n's. In the second sum change the lower limit to -1, with the understanding that $a_{-1} \equiv 0$. Then (73) becomes

$$\sum_{n=-1}^\infty \left[(n+r+1)(n+r)a_{n+1} + a_n\right] x^{n+r} = 0,$$

so we have the recursion formula

$$(n+r+1)(n+r)a_{n+1} + a_n = 0, \qquad (a_{-1} \equiv 0,\ a_0 \neq 0) \tag{74}$$

for $n = -1, 0, 1, 2, \ldots$. Setting $n = -1$, and using $a_{-1} = 0$ and $a_0 \neq 0$, gives the indicial equation

$$r(r-1) = 0, \tag{75}$$

with roots $r_1 = 1$ and $r_2 = 0$. These differ by an integer, so that the problem is of case (iii) type. Let us try the smaller root first. That root will, according to Theorem 4.3.1, lead to both solutions or to neither. With $r = r_2 = 0$, (74) becomes $(n+1)na_{n+1} + a_n = 0$. Having already used $n = -1$, to obtain (74), we next set $n = 0$. That step gives $0 + a_0 = 0$, so that $a_0 = 0$, which contradicts the assumption that $a_0 \neq 0$. Thus, $r = r_2 = 0$ yields *no* solutions. Thus, we will use the larger root, $r = r_1 = 1$, to obtain one solution, and then (41c) to obtain the other.

With $r = r_1 = 1$, (74) gives

$$a_{n+1} = -\frac{1}{(n+2)(n+1)} a_n. \tag{76}$$

Working out the first several a_n's from (76), we find that

$$a_n = -\frac{(-1)^n}{(n+1)(n!)^2} a_0,$$

so

$$y(x) = \sum_0^\infty \frac{(-1)^n a_0}{(n+1)(n!)^2} x^{n+1} = a_0 y_1(x),$$

where

$$y_1(x) = \sum_0^\infty \frac{(-1)^n}{(n+1)(n!)^2} x^{n+1}. \tag{77}$$

[Remember, throughout, that $0! = 1$ and $(-1)^0 = 1$.]

To find y_2, we use (41c) and seek

$$y_2(x) = \kappa y_1(x) \ln x + \sum_0^\infty d_n x^n. \tag{78}$$

Putting (78) into (72) gives

$$\begin{aligned} \kappa x^2 y_1'' \ln x + 2\kappa x y_1' - \kappa y_1 + \sum_0^\infty n(n-1) d_n x^n \\ + \kappa x y_1 \ln x + \sum_0^\infty d_n x^{n+1} = 0. \end{aligned} \tag{79}$$

Cancelling the $\ln x$ terms [because y_1 satisfies (72)], re-expressing the last sum in (79) as $\sum_0^\infty d_n x^{n+1} = \sum_1^\infty d_{m-1} x^m = \sum_0^\infty d_{n-1} x^n$, where $d_{-1} \equiv 0$, and putting (77) in for the y_1 terms, (79) becomes

$$\begin{aligned} \sum_0^\infty \left[n(n-1) d_n + d_{n-1} \right] x^n &= \kappa y_1 - 2\kappa x y_1' \\ &= -\kappa \sum_0^\infty \frac{(-1)^n (2n+1)}{(n+1)(n!)^2} x^{n+1} \\ &= -\kappa \sum_1^\infty \frac{(-1)^{n-1}(2n-1)}{n[(n-1)!]^2} x^n, \end{aligned} \tag{80}$$

where, to obtain the last equality, we let $n + 1 = m$ and then set $m = n$. Equating coefficients of like powers of x gives the recursion formula

$$n(n-1) d_n + d_{n-1} = -\kappa \frac{(-1)^{n-1}(2n-1)}{n[(n-1)!]^2} \tag{81}$$

for $n = 1, 2, \ldots$. [We can begin with $n = 1$ because equating the constant terms on both sides of (81) merely gives $0 = 0$.] Letting $n = 1, 2, \ldots$ gives

$$\begin{aligned} n = 1: &\quad d_0 = -\kappa, \\ n = 2: &\quad d_2 = \frac{3}{4}\kappa - \frac{1}{2} d_1, \\ n = 3: &\quad d_3 = -\frac{7}{36}\kappa + \frac{1}{12} d_1, \\ n = 4: &\quad d_4 = \frac{35}{1728}\kappa - \frac{1}{144} d_1, \end{aligned} \tag{82}$$

and so on, where d_1 remains arbitrary. Thus, the series in (78) is

$$\begin{aligned} \sum_0^\infty d_n x^n = \kappa \left(-1 + \frac{3}{4} x^2 - \frac{7}{36} x^4 + \frac{35}{1728} x^5 + \cdots \right) \\ + d_1 \left(x - \frac{1}{2} x^2 + \frac{1}{12} x^3 - \frac{1}{144} x^4 + \cdots \right). \end{aligned} \tag{83}$$

The series multiplying d_1, on the right side of (83), is identical to $y_1(x)$, given by (77), so we can set $d_1 = 0$ without loss. With $d_1 = 0$, we see that the entire right side of (78) is scaled by κ, which has remained arbitrary, so there is no loss in setting $\kappa = 1$.

Thus, $y_2(x)$ is given by (78), wherein $y_1(x)$ is given by (77) and the d_n's by (81), with d_1 taken to be zero and $\kappa = 1$. ▮

EXAMPLE 6. *Case (iii).* Solve

$$4x^2y'' + 4xy' - y = 0 \tag{84}$$

by the method of Frobenius. This has been a long and arduous section so we will only outline the solution to (84). Seeking a Frobenius expansion $y(x) = \sum_0^\infty a_n x^{n+r}$ about the regular singular point $x = 0$, we obtain the indicial equation $4r^2 - 1 = 0$, so $r = \pm 1/2$, which corresponds to case (iii) of Theorem 4.3.1. We find that the larger root $r_1 = 1/2$ leads to the one-term solution $y(x) = a_0x^{1/2}$ (i.e., $a_1 = a_2 = \cdots = 0$), and that the smaller root $r_2 = -1/2$ leads to $y(x) = a_0x^{-1/2} + a_1x^{1/2}$ (i.e., $a_2 = a_3 = \cdots = 0$), which is the general solution. We did not, in (84), specify the x interval of interest. Suppose that it is $x < 0$. Then a general solution of (84) is $y(x) = a_0|x|^{-1/2} + a_1|x|^{1/2}$, and that solution is valid on the entire interval $x < 0$.

In fact, (84) is an elementary equation, a Cauchy-Euler equation, so we could have solved it more easily. But we wanted to show that it can nonetheless be solved by the Frobenius method, and that that method does indeed give the correct one-term solutions. ▮

One final point: what if the indicial equation gives complex roots $r = \alpha \pm i\beta$? This issue came up in Section 3.6.1 as well, for the Cauchy-Euler equation. Our treatment here is virtually the same as in Section 3.6.1 and is left for Exercise 10.

Closure. The Frobenius theory, embodied in Theorem 3.4.1, enables us to find a general solution to any second-order linear ordinary differential equaton with a regular singular point at $x = 0$, in the form of generalized power series expansions about that point, possibly with $\ln x$ included. There are exactly three possible cases: if the roots of the indicial equation (38) are r_1, r_2, where $r_1 \geq r_2$ if they are real, then if the roots are distinct and not differing by an integer (which includes the case where the roots are complex) then LI solutions are given by (40) and (41a); if the roots are repeated then LI solutions are given by (40) and (41b); and if $r_1 - r_2$ is an integer then LI solutions are given by (40) and (41c). Theorem 3.4.1 is by no means of theoretical interest alone, since applications, especially the solution by separation of variables of the classical partial differential equations of mathematical physics and engineering (such as the diffusion, Laplace, and wave equations), often lead to nonconstant-coefficient second-order linear differential equations with regular singular points, such as the well known Legendre and Bessel equations. We devote Sections 4.4 and 4.6 to those two important cases.

Computer software. It is fortunate that computer-algebra systems can even generate Frobenius-type solutions, fortunate because the hand calculations can be quite tedious, as our examples have shown. Thus, we urge you to study the theory in this section on the one hand and to learn how to use such software as well. To illustrate, let us use the *Maple* dsolve command (discussed at the end of Section 4.2) to obtain a Frobenius-type solution of the differential equation $xy''+y=0$ about the regular singular point $x=0$; this was our Example 6. Enter

$$\text{dsolve}(x * \ \text{diff}(y(x),\ x,\ x) + y(x) = 0,\ y(x),\ \text{type} = \text{series});$$

and return. The resulting output

$$\begin{aligned} y(x) = \_C1\,x &\left(1 - \frac{1}{2}x + \frac{1}{12}x^2 - \frac{1}{144}x^3 + \frac{1}{2880}x^4 - \frac{1}{86400}x^5 + O(x^6)\right) \\ +\_C2 &\left[\ln(x)\left(-x + \frac{1}{2}x^2 - \frac{1}{12}x^3 + \frac{1}{144}x^4 - \frac{1}{2880}x^5 + O(x^6)\right)\right. \\ &\left. + \left(1 - \frac{3}{4}x^2 + \frac{7}{36}x^3 - \frac{35}{1728}x^4 + \frac{101}{86400}x^5 + O(x^6)\right)\right] \end{aligned}$$

is found to agree with the general solution that we generated in Example 5.

EXERCISES 4.3

1. For each equation, identify all singular points (if any), and classify each as regular or irregular. For each regular singular point use Theorem 4.3.1 to determine the minimum possible radii of convergence of the series that will result in (40) and (41) (but you need not work out those series).

(a) $y'' - x^3y' + xy = 0$
(b) $xy'' - (\cos x)y' + 5y = 0$
(c) $(x^2-3)y'' - y = 0$
(d) $x(x^2+3)y'' + y = 0$
(e) $(x+1)^2y'' - 4y' + (x+1)y = 0$
(f) $y'' + (\ln x)y' + 2y = 0$
(g) $(x-1)(x+3)^2y'' + y' + y = 0$
(h) $xy'' + (\sin x)y' - (\cos x)y = 0$
(i) $x(x^4+2)y'' + y = 0$
(j) $(x^4-1)y'' + xy' - x^2y = 0$
(k) $(x^4-1)^3y'' + (x^2-1)^2y' - y = 0$
(l) $(x^4-1)^3y'' - 3(x+1)^2y' + x(x+1)y = 0$
(m) $(xy')' - 5y = 0$
(n) $\left[x^3(x-1)y'\right]' + 2y = 0$
(o) $2x^2y'' - xy' + 7y = 0$
(p) $xy'' + 4y' = 0$
(q) $x^2y'' - 3y = 0$
(r) $2x^2y'' + \sqrt{\pi}\,y = 0$

2. Sometimes one can change an irregular singular point to a regular singular point, by suitable change of variables, so that the Frobenius theory can be applied. The purpose of this exercise is to present such a case. We noted, in Example 3, that $y'' + \sqrt{x}\,y = 0$ $(x > 0)$ has an irregular singular point at $x = 0$, because of the $\sqrt{x}$.

(a) Show that if we change the independent variable from x to t, say, according to $\sqrt{x} = t$, then the equation on $y(x(t)) = Y(t)$ is

$$Y''(t) - \frac{1}{t}Y'(t) + 4t^3Y(t) = 0. \qquad (t > 0) \qquad (2.1)$$

(b) Show that (2.1) has a regular singular point at $t = 0$ (which point corresponds to $x = 0$).
(c) Obtain a general solution of (2.1) by the Frobenius method. (If possible, give the general term of any series obtained.) Putting $t = \sqrt{x}$ in that result, obtain the corresponding general solution of $y'' + \sqrt{x}\,y = 0$. Is that general solution for $y(x)$ of Frobenius form? Explain.
(d) Use computer software to find a general solution.

3. In each case, there is a regular singular point at the left end of the stated x interval; call that point x_0. Merely introduce a change of independent variable, from x to t, according to $x - x_0 = t$, and obtain the new differential equation on $y(x(t)) = Y(t)$. You need not solve that equation.

(a) $(x-1)y'' + y' - y = 0, \qquad (1 < x < \infty)$
(b) $(x^2-1)y'' + y = 0, \qquad (1 < x < \infty)$
(c) $(x+3)y'' - 2(x+3)y' - 4y = 0, \qquad (-3 < x < \infty)$
(d) $(x-5)^2y'' + 2(x-5)y' - y = 0, \qquad (5 < x < \infty)$

4. Derive the series solution (25).

5. Make up a differential equation that will have as the roots of its indicial equation

(a) $1, 4$ (b) $3, 3$ (c) $1/2, 2$ (d) $-1/2, 1/2$
(e) $2 \pm 3i$ (f) $-1, -1$ (g) $-2/3, 5$ (h) $-1 \pm i$
(i) $(1 \pm 2i)/3$ (j) $5/4, 8/3$

6. In each case verify that $x = 0$ is a regular singular point, and use the method of Frobenius to obtain a general solution $y(x) = Ay_1(x) + By_2(x)$ of the given differential equation, on the interval $0 < x < \infty$. That is, determine $y_1(x)$ and $y_2(x)$. On what interval can you be certain that your solution is valid? HINT: See Theorem 4.3.1.

(a) $2xy'' + y' + x^3y = 0$
(b) $xy'' + y' - xy = 0$
(c) $xy'' + y' + x^8y = 0$
(d) $xy'' + y' + xy = 0$
(e) $x^2y'' + xy' - y = 0$
(f) $x^2y'' - x^2y' - 2y = 0$
(g) $x^2y'' + xy' - (1+2x)y = 0$
(h) $x^2y'' + xy' - y = 0$
(i) $xy'' + xy' + (1+x)y = 0$
(j) $3xy'' + y' + y = 0$
(k) $x(1+x)y'' + y = 0$
(l) $x^2(2+x)y'' - y = 0$
(m) $x^2y'' - (2+3x)y = 0$
(n) $5xy'' + y' + 8x^2y = 0$
(o) $xy'' + e^xy = 0$
(p) $2xy'' + e^xy' + y = 0$
(q) $16x^2y'' + 8xy' - 3y = 0$
(r) $16x^2y'' + 8xy' - (3+x)y = 0$
(s) $x^2y'' + xy' + (\sin x)y = 0$
(t) $5(xy)'' - 9y' + xy = 0$
(u) $(xy')' - y = 0$
(v) $(xy')' - 2y' - y = 0$

7. (a)–(x) Use computer software to obtain a general solution of the corresponding differential equation in Exercise 6.

8. Use the method of Frobenius to obtain a general solution to the equation $xy'' + cy' = 0$ on $x > 0$, where c is a real constant. You may need to treat different cases, depending upon c.

9. (a) The equation

$$(x^2 - x)y'' + (4x-2)y' + 2y = 0, \quad (0 < x < 1) \tag{9.1}$$

has been "rigged" to have, as solutions, $1/x$ and $1/(1-x)$. Solve (9.1) by the method of Frobenius, and show that you do indeed obtain those two solutions.
(b) You may have wondered how we made up the equation (9.1) so as to have the two desired solutions. Here, we ask you to make up a linear homogenous second-order differential equation that has two prescribed LI solutions $F(x)$ and $G(x)$.

10. (*Complex roots*) Since $p(x)$ and $q(x)$ are real-valued functions, p_0 and q_0 are real. Thus, if the indicial equation (38) has complex roots they will be complex conjugates, $r = \alpha \pm i\beta$, so case (i) of Theorem 3.4.1 applies, and the method of Frobenius will give a general solution of the form

$$\begin{aligned} y(x) &= Ay_1(x) + By_2(x) \\ &= Ax^{\alpha+i\beta}\textstyle\sum_0^\infty a_nx^n + Bx^{\alpha-i\beta}\sum_0^\infty b_nx^n. \end{aligned} \tag{10.1}$$

(a) Show that the b_n's will be the complex conjugates of the a_n's: $b_n = \bar{a}_n$.
(b) Recalling, from Section 3.6.1, that

$$x^{\alpha \pm i\beta} = x^\alpha \left[\cos(\beta \ln x) \pm i \sin(\beta \ln x)\right], \tag{10.2}$$

show that (10.1) [with b_n replaced by $\bar{a}_n$, according to the result found in part (a) above] can be re-expressed in terms of real functions as

$$\begin{aligned} y(x) = Cx^\alpha &\left[\cos(\beta \ln x)\textstyle\sum_0^\infty c_nx^n \right. \\ &\left. - \sin(\beta \ln x)\textstyle\sum_0^\infty d_nx^n\right] \\ +Dx^\alpha &\left[\cos(\beta \ln x)\textstyle\sum_0^\infty d_nx^n \right. \\ &\left. + \sin(\beta \ln x)\textstyle\sum_0^\infty c_nx^n\right], \end{aligned} \tag{10.3}$$

where c_n, d_n are the real and imaginary parts of a_n, respectively: $a_n = c_n + id_n$.
(c) Find a general solution of the form (10.3) for the equation

$$x^2y'' + x(1+x)y' + y = 0.$$

That is, determine α, β and c_n, d_n in (10.3), through $n = 3$, say.
(d) The same as (c), for $x^2y'' + xy' + (1-x)y = 0$.

4.4 Legendre Functions

4.4.1. Legendre polynomials. The differential equation

$$\boxed{\left(1-x^2\right)y''-2xy'+\lambda y=0,} \tag{1}$$

where λ is a constant, is known as **Legendre's equation**, after the French mathematician *Adrien-Marie Legendre* (1752–1833). The x interval of interest is $-1<x<1$, and (1) has regular singular points at each endpoint, $x=\pm 1$. In this section we study aspects of the Legendre equation and its solutions that will be needed in applications in later chapters. There, we will be interested in power series solutions about the ordinary point $x=0$,

$$y(x)=\sum_{k=0}^{\infty} a_k x^k. \tag{2}$$

Putting (2) into (1) leads to the recursion formula (Exercise 1)

$$a_{k+2}=\frac{k(k+1)-\lambda}{(k+1)(k+2)}a_k. \qquad (k=0,1,2,\cdots) \tag{3}$$

Setting $k=0,1,2,\ldots$, in turn, shows that a_0 and a_1 are arbitrary, and that subsequent a_k's can be expressed, alternately, in terms of a_0 and a_1:

$$a_2=-\frac{\lambda}{2}a_0, \quad a_3=\frac{2-\lambda}{6}a_1, \quad a_4=-\frac{(6-\lambda)\lambda}{24}a_0,$$

and so on, and we have the general solution

$$\begin{aligned} y(x)&=a_0\left[1-\frac{\lambda}{2}x^2-\frac{(6-\lambda)\lambda}{24}x^4-\frac{(20-\lambda)(6-\lambda)\lambda}{720}x^6-\cdots\right] \\ &\quad +a_1\left[x+\frac{2-\lambda}{6}x^3+\frac{(12-\lambda)(2-\lambda)}{120}x^5+\cdots\right] \\ &=a_0y_1(x)+a_1y_2(x) \end{aligned} \tag{4}$$

of (1). To determine the radii of convergence of the two series in (4) we can use the recursion formula (3) and Theorem 4.2.2, provided that we realize that the a_{k+2} on the left side of (3) is really the next coefficient after a_k, the "a_{k+1}" in Theorem 4.2.2 since every other term in each series is missing. Thus, (3) gives

$$\lim_{k\to\infty}\left|\frac{\text{“}a_{k+1}\text{”}}{a_k}\right|=\lim_{k\to\infty}\left|\frac{k(k+1)-\lambda}{(k+1)(k+2)}\right|=1, \tag{5}$$

and it follows from Theorem 4.2.2 that $R=1$, so each series converges in $-1<x<1$.

In physical applications of the Legendre equation, such as finding the steady-state temperature distribution within a sphere subjected to a known temperature distribution on its surface, one needs solutions that are *bounded* on $-1 < x < 1$. [$F(x)$ being **bounded** on an interval I means that there exists a finite constant M such that $|F(x)| \leq M$ for all x in I. If $F(x)$ is not bounded then it is **unbounded**.] However, for arbitrary values of the parameter λ the functions $y_1(x)$ and $y_2(x)$ given in (4) grow unboundedly as $x \to \pm 1$, as illustrated in Fig. 1 for $\lambda = 1$. If you studied Section 4.3, then you could investigate why that is so by developing a Frobenius-type solution about the right endpoint $x = 1$, which is a regular singular point of (1). Doing so, you would find a $\ln(1-x)$ term, within the solutions, which is infinite at $x = 1$. Similarly, a Frobenius solution about $x = -1$ would reveal a $\ln(1+x)$ term, which is infinite at $x = -1$. Evidently, $y_1(x)$ and $y_2(x)$, above, contain linear combinations of $\ln(1-x)$ and $\ln(1+x)$ [of course, one cannot see them explicitly in (4) because (4) is an expansion about $x = 0$, not $x = 1$ or $x = -1$] so they grow unboundedly as $x \to \pm 1$.

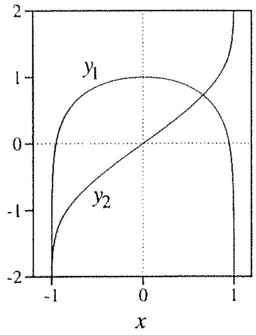


Figure 1. $y_1(x)$ and $y_2(x)$ in (4), for $\lambda = 1$.

Nonetheless, for certain specific values of λ one series or the other, in (4), will *terminate* and thereby be bounded on the interval since it is then a finite degree polynomial! Specifically, if λ happens to be such that

$$\lambda = n(n+1) \tag{6}$$

for any integer $n = 0, 1, 2, \ldots$, then we can see from (3) that one of the two series terminates at $k = n$: if $\lambda = n(n+1)$, where n is even, then the even-powered series terminates at $k = n$ (because $a_{n+2} = a_{n+4} = \cdots = 0$). For example, if $n = 2$ and $\lambda = 2(2+1) = 6$, then the $6 - \lambda$ factor in every term after the second, in the even-powered series, causes all those terms to vanish, so the series terminates as $1 - 3x^2$. Similarly, if $\lambda = n(n+1)$, where n is odd, then the odd-powered series terminates at $k = n$. The first five such λ's, and their corresponding polynomial solutions of (1), are shown in the second and third columns of Table 1. These

Table 1. The first five Legendre polynomials.

n	$\lambda = n(n+1)$	Polynomial Solution	Legendre Polynomial $P_n(x)$
0	0	1	$P_0(x) = 1$
1	2	x	$P_1(x) = x$
2	6	$1 - 3x^2$	$P_2(x) = \frac{1}{2}(3x^2 - 1)$
3	12	$x - \frac{5}{3}x^3$	$P_3(x) = \frac{1}{2}(5x^3 - 3x)$
4	20	$1 - 10x^2 + \frac{35}{3}x^4$	$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3)$

polynomial solutions can, of course, be scaled by any desired numerical factor. Scaling them to equal unity at $x = 1$, by convention, they become the so-called **Legendre polynomials**. Thus, the Legendre polynomial $P_n(x)$ is a polynomial solution of the Legendre equation

$$(1-x^2)y'' - 2xy' + n(n+1)y = 0, \tag{7}$$

scaled so that $P_n(1) = 1$. In fact, it can be shown that they are given explicitly by the formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}\left[(x^2-1)^n\right], \tag{8}$$

which result is known as **Rodrigues's formula**.

4.4.2. Orthogonality of the P_n's. For reasons that will not be evident until we study vector spaces (Section 9.6), the integral formula

$$\boxed{\int_{-1}^{1} P_j(x)P_k(x)\,dx = 0, \qquad (j \neq k)} \tag{9}$$

is known as the **orthogonality relation**. By virtue of (9), we say that $P_j(x)$ and $P_k(x)$ are orthogonal to each other – provided that they are different Legendre polynomials (i.e., $j \neq k$).

Proof of (9) is not difficult. Noting that the $(1-x^2)y'' - 2xy'$ terms in (7) can be combined as $[(1-x^2)y']'$, we begin by considering $\int_{-1}^{1}\left[(1-x^2)P_j'\right]' P_k\,dx$ and integrating by parts until all the derivatives have been transferred from P_j to P_k:

$$\begin{aligned}\int_{-1}^{1}\left[(1-x^2)P_j'\right]' P_k\,dx &= (1-x^2)\,P_j'P_k\big|_{-1}^{1} - \int_{-1}^{1}(1-x^2)P_j'P_k'\,dx \\ &= 0 - (1-x^2)\,P_k'P_j\big|_{-1}^{1} + \int_{-1}^{1} P_j\left[(1-x^2)P_k'\right]'\,dx. \end{aligned}\tag{10}$$

The next to last term is zero because of the $1-x^2$ factor, just as the boundary term following the first equal sign is zero. Since P_j and P_k are solutions of the Legendre equation

$$\left[(1-x^2)y'\right]' + n(n+1)y = 0 \tag{11}$$

for $n = j$ and k, respectively, we can use (11) to re-express (10) as

$$-j(j+1)\int_{-1}^{1} P_jP_k\,dx = -k(k+1)\int_{-1}^{1} P_jP_k\,dx \tag{12}$$

or

$$[k(k+1) - j(j+1)]\int_{-1}^{1} P_j(x)P_k(x)\,dx = 0. \tag{13}$$

Since $j \neq k$, it follows from (13) that $\int_{-1}^{1} P_jP_k\,dx = 0$, as was to be proved.

We will see later that (9) is but a special case of the more general orthogonality relation found in the Sturm-Liouville theory, which theory will be essential to us when we solve partial differential equations.

4.4.3. Generating function and properties. Besides (9), another important property of Legendre polynomials is expressed by the formula

$$\boxed{\left(1-2xr+r^2\right)^{-1/2}=\sum_0^\infty P_n(x)r^n. \qquad (|x|\le 1,\ |r|<1)} \tag{14}$$

That is, if we regard the left side of (14) as a function of r and expand it in a Taylor series about $r=0$, then the coefficient of r^n turns out to be $P_n(x)$. Thus, $\left(1-2xr+r^2\right)^{-1/2}$ is called the **generating function** for the P_n's (Exercise 4).

Equation (14) is the source of considerable additional information about the P_n's. For instance, by changing x to $-x$ in (14) it can be seen that

$$P_n(-x)=(-1)^n P_n(x). \tag{15}$$

Now, if $f(-x)=f(x)$, then the graph of f is symmetric about $x=0$ and we say that f is an **even function** of x. If, instead, $f(-x)=-f(x)$, then the graph of f is antisymmetric about $x=0$ and we say that f is an **odd function** of x. Noting that the $(-1)^n$ is $+1$ if n is an even integer and -1 if n is an odd integer, then we see from (15) that $P_n(x)$ is an even function of x if n is an even integer, and an odd function of x if n is an odd integer, as is seen to be true for the P_n's that are shown in Fig. 2.

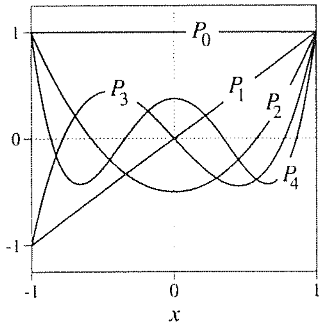


Figure 2. Graphs of the first five Legendre polynomials.

Also, by taking $\partial/\partial r$ of (14) one can show (Exercise 6) that

$$nP_n(x)=(2n-1)xP_{n-1}(x)-(n-1)P_{n-2}(x), \qquad (n=2,3,\dots) \tag{16}$$

which is a recursion relation giving P_n in terms of P_{n-1} and P_{n-2}. Or by taking $\partial/\partial x$ of (14) instead, one can show (Exercise 7) that

$$P_n'(x)-2xP_{n-1}'(x)+P_{n-2}'(x)=P_{n-1}(x). \qquad (n=2,3,\dots) \tag{17}$$

Finally, squaring both sides of (14) and integrating on x from -1 to $+1$, and using the orthogonality relation (9), one can show that

$$\boxed{\int_{-1}^{1}\left[P_n(x)\right]^2 dx=\frac{2}{2n+1}, \qquad (n=0,1,2,\dots)} \tag{18}$$

which result is a companion to (9); it covers the case where $j=k$ ($=n$, say). We will need (9) and (18) in later chapters.

Closure. Our principal application of Legendre's equation and Legendre polynomials, in this text, is in connection with the solution of the Laplace equation in spherical coordinates. There, we need to know how to expand a given function in terms of the Legendre polynomials $P_0(x), P_1(x), P_2(x), \dots$, and the theory behind such expansions is covered in Section 17.6 on the Sturm-Liouville theory.

To help put that idea into perspective, recall from a first course in physics or mechanics or calculus that one can expand any given vector in 3-space in terms of orthogonal (perpendicular) vectors "$\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$." That fact is of great importance and was probably used extensively in those courses. Remarkably, we will be able to generalize the idea of vectors so as to regard *functions* as vectors. It will turn out that the set of Legendre polynomials $P_0, P_1, \dots$ constitute an infinite orthogonal set of vectors such that virtually any given function defined on $-1 \le x \le 1$ can be expanded in terms of them, just as any given "arrow vector" in 3-space can be expanded in terms of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. In the present section we have not gotten that far, but the results obtained here will be used later, when we finish the story.

For a more extensive treatment of Legendre functions, Bessel functions, and the various other important special functions of mathematical physics, see, for instance, D. E. Johnson and J. R. Johnson, *Mathematical Methods in Engineering and Physics* (Englewood Cliffs, NJ: Prentice Hall, 1982). Even short of a careful study of the other special functions – such as those associated with the names Bessel, Hermite, Laguerre, Chebyshev, and Mathieu – we recommend browsing through a book like Johnson and Johnson so as to at least become aware of these functions and the circumstances under which they arise.

Computer software. In *Maple*, $P_n(x)$ is denoted as $P(n, x)$. To obtain $P_7(x)$, say, enter

```
with(orthopoly):
```

and return, then

$$P(7, x);$$

and return. The result is $\frac{429}{16}x^7 - \frac{693}{16}x^5 + \frac{315}{16}x^3 - \frac{35}{16}x$.

EXERCISES 4.4

1. Putting (2) into (1), derive the recursion formula (3).

2. Obtain (4) using computer software.

3. Use Rodrigues's formula, (8), to reproduce the first five Legendre polynomials, cited in Table 1.

4. Expanding the left-hand side of (14) in a Taylor series in r, about $r = 0$, through r^3, say, verify that the coefficients of $r^0, \dots, r^3$ are indeed $P_0(x), \dots, P_3(x)$, respectively.

5. We stated that by changing x to $-x$ in (14) it can be seen that $P_n(-x) = (-1)^n P_n(x)$. Show those steps and explain your reasoning.

6. (a) We stated that by taking $\partial/\partial r$ of (14) one obtains (16). Show those steps and explain your reasoning.
(b) Verify (16) for $n = 2$.
(c) Verify (16) for $n = 3$.

7. (a) We stated that by taking $\partial/\partial x$ of (14) one obtains (17). Show those steps and explain your reasoning.
(b) Verify (17) for $n = 2$.
(c) Verify (17) for $n = 3$.

8. (a) Derive (18) as follows. Squaring (14) and integrating from -1 to 1, obtain

$$\int_{-1}^{1} \frac{dx}{1 - 2rx + r^2} = \int_{-1}^{1} \sum_{m=0}^{\infty} r^m P_m(x) \sum_{n=0}^{\infty} r^n P_n(x)\, dx. \tag{8.1}$$

Integrating the left side, and using the orthogonality relation (9) to simplify the right side, obtain

$$\frac{1}{r} \ln\left(\frac{1+r}{1-r}\right) = \sum_{n=0}^{\infty} \left\{ \int_{-1}^{1} [P_n(x)]^2\, dx \right\} r^{2n}. \tag{8.2}$$

Finally, expanding the left-hand side in a Taylor series in r, show that (18) follows.
(b) Verify (18), by working out the integral, for the cases $n = 0, 1,$ and 2.

9. (*Integral representation of* P_n) It can be shown that

$$P_n(x) = \tfrac{1}{\pi}\int_0^\pi \left(x + \sqrt{x^2-1}\,\cos t\right)^n dt, \qquad (n = 0, 1, 2, \ldots) \tag{9.1}$$

which is called **Laplace's integral form** for $P_n(x)$. Here, we ask you to verify (9.1) for the cases $n = 0, 1$, and 2, by working out the integral for those cases.

10. We sought power series expansions of (1) about the ordinary point $x = 0$ and, for the case where $\lambda = n(n+1)$, we obtained a bounded solution [namely, the Legendre polynomial $P_n(x)$] and an unbounded solution. Instead, seek a Frobenius-type solution about the regular singular points $x = 1$ and $x = -1$, for the case where

(a) $n = 0 \quad (\lambda = 0)$
(b) $n = 1 \quad (\lambda = 1)$
(c) $n = 2 \quad (\lambda = 2)$

11. (*Legendre functions of second kind*) For the Legendre equation (7) on the interval $-1 < x < 1$, we obtained the bounded solution $y(x) = P_n(x)$. In this exercise we seek a second LI solution, denoted as $Q_n(x)$ and called the **Legendre function of the second kind**. Then the general solution of (7) can be expressed as

$$y(x) = AP_n(x) + BQ_n(x). \tag{11.1}$$

(a) For the special case $n = 0$, solve (7) and show that a second LI solution is $\ln[(1+x)/(1-x)]$. Scaling this solution by 1/2, we define

$$Q_0(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right). \tag{11.2}$$

Sketch the graph of $Q_0(x)$ on $-1 < x < 1$, and notice that $|Q_0(x)| \to \infty$ as $x \to \pm 1$.
(b) More generally, consider any nonnegative integer n. With only $P_n(x)$ in hand, seek a second solution (by reduction of order) in the form $y(x) = A(x)P_n(x)$, and show that $Q_n(x)$ is given by

$$Q_n(x) = C_n P_n(x) \int^x \frac{dt}{(1-t^2)\,[P_n(t)]^2} + D_n P_n(x). \tag{11.3}$$

(c) Evaluating the integral in (11.3), show that the first two Q_n's are

$$Q_0(x) = \frac{1}{2}C_0 \ln\left(\frac{1+x}{1-x}\right) + D_0, \tag{11.4a}$$

$$Q_1(x) = C_1\left[\frac{x}{2}\ln\left(\frac{1+x}{1-x}\right) - 1\right] + D_1 x. \tag{11.4b}$$

By convention, choose $C_0 = 1$, $D_0 = 0$, $C_1 = 1$ and $D_1 = 0$, so that

$$Q_0(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right), \tag{11.5a}$$

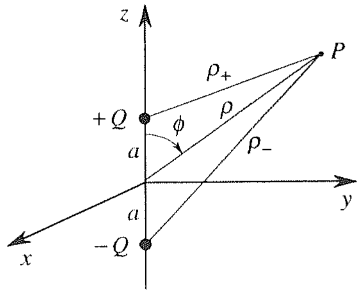
$$Q_1(x) = \frac{x}{2}\ln\left(\frac{1+x}{1-x}\right) - 1. \tag{11.5b}$$

(d) The recursion formula (16) holds for the Q_n's as well as the P_n's. Thus, with Q_0 and Q_1 in hand we can use (16) to obtain Q_2, Q_3, and so on. Do that: show that

$$Q_2(x) = \frac{3x^2-1}{4}\ln\left(\frac{1+x}{1-x}\right) - \frac{3}{2}x, \tag{11.6}$$

and obtain $Q_3(x)$ as well.

12. (*Electric field induced by two charges*) Given a positive charge Q and a negative charge $-Q$, a distance $2a$ apart, let us introduce a coordinate system as in the figure below. Since the charges lie on the z axis, it follows that the electric field that they induce will be symmetric about the z axis.



(a) Specifically, the *electric potential* (i.e., the *voltage*) Φ induced by a charge q is $\Phi = (1/4\pi\epsilon_0)(q/r)$, where the physical constant ϵ_0 is the permittivity of free space and r is the distance from the charge to the field point. Thus the potential induced at the field point P shown in the figure is

$$\Phi = \frac{1}{4\pi\epsilon_0}\left(\frac{Q}{\rho_+} - \frac{Q}{\rho_-}\right). \tag{12.1}$$

Show that (12.1) gives

$$\Phi(\rho,\phi)=\frac{Q}{4\pi\epsilon_0}\left(\frac{1}{\sqrt{a^2+\rho^2-2a\rho\cos\phi}}-\frac{1}{\sqrt{a^2+\rho^2+2a\rho\cos\phi}}\right),$$

$$=\begin{cases}\dfrac{1}{4\pi\epsilon_0}\dfrac{2Q}{\rho}\displaystyle\sum_{n=1,3,\dots}^{\infty}\left(\frac{a}{\rho}\right)^n P_n(\cos\phi) & (\rho>a)\\ \dfrac{1}{4\pi\epsilon_0}\dfrac{2Q}{a}\displaystyle\sum_{n=1,3,\dots}^{\infty}\left(\frac{\rho}{a}\right)^n P_n(\cos\phi). & (\rho<a)\end{cases} \tag{12.2}$$

(b) With the point P fixed, imagine letting a become arbitrarily small. Show, from (12.2), that we obtain

$$\Phi(\rho,\phi)\sim\frac{1}{4\pi\epsilon_0}2Qa\frac{\cos\phi}{\rho^2} \tag{12.3}$$

as $a\to 0$. Thus, $\Phi(\rho,\phi)\to 0$ as $a\to 0$, as makes sense, because the positive and negative charges cancel each other as they are moved together. However, observe that if, as a is decreased, Q is increased such that the product Qa is held constant, then (12.3) becomes

$$\Phi(\rho,\phi)\sim\frac{\mu}{4\pi\epsilon_0}\frac{\cos\phi}{\rho^2}, \tag{12.4}$$

where $\mu=2Qa$ is called the **dipole moment**, and the charge configuration is said to constitute an **electric dipole**. If, for instance, a molecule is comprised of equal and opposite charges, $+Q$ and $-Q$, displaced by a very small distance $2a$, then, even though a is not tending to zero and Q to infinity, the field induced by that molecule is approximately equal to that of an idealized dipole of strength $\mu=2Qa$, at points sufficiently far away (i.e., for $\rho/a >> 1$).

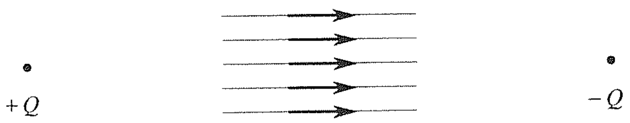
(c) As a different limit of interest, imagine the point P as fixed, and this time let a become arbitrarily *large*. Show, from (12.2), that we obtain

$$\Phi(\rho,\phi)\sim\frac{1}{4\pi\epsilon_0}\frac{2Q}{a^2}\rho\cos\phi=\frac{1}{4\pi\epsilon_0}\frac{2Q}{a^2}z \tag{12.5}$$

as $a\to\infty$. Notice that if, as a is increased, Q is increased such that Q/a^2 is held constant, then the *electric field intensity* E (which, we recall from a course in physics, is the negative of the derivative of the potential) is a constant:

$$E=-\frac{d\Phi}{dz}=-\frac{1}{4\pi\epsilon_0}\frac{Q}{2a^2}; \tag{12.6}$$

that is, we have a uniform field. Thus, a uniform electric field can be thought of as resulting from moving apart two charges, $+Q$ and $-Q$, and at the same time increasing their strength Q such that Q/a^2 is held constant as $a\to\infty$. Similarly, in fluid mechanics, a uniform fluid velocity field can be thought of as resulting from moving a fluid "source" of strength $+Q$ and a fluid "sink" of strength $-Q$ apart in such a way that Q/a^2 is held constant as $a\to\infty$, where $2a$ is their separation distance, as sketched schematically in the figure.



4.5 Singular Integrals; Gamma Function

This section amounts to a diversion from the main stream of this chapter because when we come to Bessel functions in the next section we need to know about the gamma function, and to study the gamma function we need to know about singular integrals. Furthermore, both of these topics are needed in Chapter 5 on the Laplace transform. So let us take time out to introduce them here.

4.5.1. Singular integrals. An integral is said to be **singular** (or **improper**) if one or both integration limits are infinite and/or if the integrand is unbounded on the interval; otherwise, it is **regular** (or **proper**).

For example, if

$$\begin{aligned} I_1 &= \textstyle\int_3^\infty xe^{-2x}\,dx, & I_2 &= \textstyle\int_0^5 e^{-x}\,dx/\sqrt{x}, \\ I_3 &= \textstyle\int_{-1}^2 dx/(x-1), & I_4 &= \textstyle\int_0^{100} \sqrt{x}\,e^x\,dx, \end{aligned} \tag{1}$$

then I_1 is singular due to the infinite limit, I_2 is singular because the integrand tends to ∞ as $x \to 0$, I_3 is singular because the integrand is unbounded (tends to $-\infty$ as $x \to 1$ from the left, and tends to $+\infty$ as $x \to 1$ from the right), and I_4 is regular.

Most of our interest will be in integrals that are singular by virtue of an infinite upper limit (illustrated by I_1) and/or a singularity in the integrand at the left endpoint (illustrated by I_2), so we limit this brief discussion to those cases. Other cases are considered in the exercises.

Consider the first type, $\int_a^\infty f(x)\,dx$. Analogous to our definition

$$\sum_{n=0}^{\infty} a_n \equiv \lim_{N\to\infty} \sum_{n=0}^{N} a_n \tag{2}$$

of an infinite series, we define

$$\boxed{I = \int_a^\infty f(x)\,dx \equiv \lim_{X\to\infty} \int_a^X f(x)\,dx.} \tag{3}$$

If the limit exists, we say that I is **convergent**; if not, it is **divergent**.

Recall, from our review of infinite series in Section 4.2, that the necessary and sufficient condition for the convergence of an infinite series is given by the Cauchy convergence theorem, but that theorem is difficult to apply. Thus, in the calculus, we studied a wide variety of specialized but more easily applied methods and theorems. For instance, one proves, in the calculus, that the p-**series**,

$$\sum_1^\infty \frac{1}{n^p}, \tag{4}$$

converges if $p > 1$ and diverges if $p \le 1$, the case $p = 1$ giving the well known (and divergent) **harmonic series** $\sum_1^\infty \frac{1}{n}$. That is, the terms need to die out fast enough, as n increases, for the series to converge. As p is increased, they die out faster and faster, and the borderline case is $p = 1$, with convergence requiring $p > 1$.

Then one establishes one or more *comparison tests*. For instance: If $S_1 = \sum_0^\infty a_n$ and $S_2 = \sum_0^\infty b_n$ are series of finite positive terms, and $a_n \sim Kb_n$ as $n \to \infty$ for some finite constant K, then S_1 and S_2 both converge or both diverge. (The lower limits are inconsequential insofar as convergence/divergence is concerned and have been taken to be 0 merely for definiteness.)

For instance, to determine the convergence or divergence of the series $S = \displaystyle\sum_1^\infty \frac{2n+3}{n^4+5}$, we observe that $\dfrac{2n+3}{n^4+5} \sim \dfrac{2}{n^3}$ as $n \to \infty$. Now, $\displaystyle\sum_1^\infty \frac{1}{n^3}$ is convergent because it is a p-series with $p = 3 > 1$, and by the comparison test stated above it follows that S is convergent too.

Our development for determining the convergence/divergence of singular integrals is analogous to the development described above for infinite series. Analogous to the p-series, we study the **horizontal p-integral**,

$$I = \int_a^\infty \frac{1}{x^p}\,dx, \qquad (a > 0) \tag{5}$$

where p is a constant. (The name "horizontal p-integral" is not standard, and is explained below.) The latter integral is simple enough so that we can determine its convergence/divergence by direct evaluation. Then we can use that result, in conjunction with comparison tests, to determine the convergence/divergence of more complicated integrals. Proceeding,

$$I = \int_a^\infty \frac{1}{x^p}\,dx = \lim_{X\to\infty}\int_a^X \frac{1}{x^p}\,dx = \begin{cases} \lim_{X\to\infty} \frac{1}{1-p}\, x^{1-p}\big|_a^X & (p \neq 1) \\ \lim_{X\to\infty} \ln x\big|_a^X . & (p = 1) \end{cases} \tag{6}$$

Now, $\lim_{X\to\infty} \ln X$ is infinite and hence does not exist, and similarly for $\lim_{X\to\infty} X^{1-p}$ if $p < 1$, whereas the latter does exist if $p > 1$. Thus,

THEOREM 4.5.1 *Horizontal p-Integral*
The horizontal p-integral, (5), converges if $p > 1$ and diverges if $p \leq 1$.

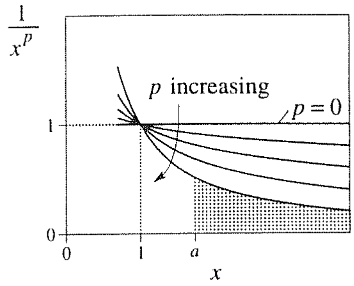


Figure 1. The effect, on $1/x^p$, of varying p.

That result is easy to remember since the p-series, likewise, converges if $p > 1$ and diverges if $p \leq 1$. Graphically, the idea is that p needs to be positive enough (namely, $p > 1$) so that the infinitely long sliver of area (shaded in Fig. 1) is squeezed thin enough to have a finite area.

We state the following comparison tests without proof.

THEOREM 4.5.2 *Comparison Tests*
Let $I_1 = \int_a^\infty f(x)\,dx$ and $I_2 = \int_a^\infty g(x)\,dx$, where $f(x)$ and $g(x)$ are positive (and bounded) on $a \leq x < \infty$.
(a) If there exist constants K and X such that $f(x) \leq Kg(x)$ for all $x \geq X$, then the convergence of I_2 implies the convergence of I_1, and the divergence of I_1 implies the divergence of I_2.
(b) If $f(x) \sim Cg(x)$ as $x \to \infty$, for some finite constant C, then I_1 and I_2 both converge or both diverge.

Of course, K must be finite. Actually, (b) is implied by (a), but we have included it explicitly since it is a simpler statement and is easier to use. Note that C cannot be zero because the notation $f(x) \sim 0$ makes no sense. That is, $f(x) \sim g(x)$ as $x \to x_0$ means that $f(x)/g(x) \to 1$ as $x \to x_0$, and $f(x)/0$ cannot possibly tend to 1.

EXAMPLE 1. Consider $I = \int_9^\infty \frac{2x+3}{x^4+5}\,dx$. Since $\frac{2x+3}{x^4+5} \sim \frac{2}{x^3}$ as $x \to \infty$, and $\int_9^\infty dx/x^3$ is a convergent p-integral ($p = 4 > 1$), it follows from Theorem 4.5.2(b) that I is convergent.

COMMENT. If, instead, the integrand were $(2x+3)/(x^2+5)$, then the integral would be divergent because $(2x+3)/(x^2+5) \sim 2/x$, and $\int_9^\infty dx/x$ is a divergent p-integral ($p = 1$). It would be incorrect to argue that the integral converges because $(2x+3)/(x^2+5) \to 0$ as $x \to \infty$. Tending to zero is not enough; the integrand must tend to zero fast enough. ▮

Since the integrand of the integral in question might not be positive, as was assumed in Theorem 4.5.2, the following theorem is useful.

THEOREM 4.5.3 *Absolute Convergence*
If $\int_a^\infty |f(x)|\,dx$ converges, then so does $\int_a^\infty f(x)\,dx$, and we say that the latter converges **absolutely**.

EXAMPLE 2. Consider $I = \int_2^\infty \frac{\sin x}{3x^2+1}\,dx$, the integrand of which is not everywhere positive. We have

$$\left|\frac{\sin x}{3x^2+1}\right| \le \frac{1}{3x^2+1} \sim \frac{1}{3x^2} \quad \text{as } x \to \infty. \tag{7}$$

Now, $\int_2^\infty dx/x^2$ is a convergent p-integral ($p = 2 > 1$). Thus, by the asymptotic relation in (7) and Theorem 4.5.2(b), $\int_2^\infty dx/(3x^2+1)$ converges. Next, by the inequality in (7) and Theorem 4.5.2(a), $\int_2^\infty \left|\sin x/(3x^2+1)\right|\,dx$ converges. Finally, by Theorem 4.5.3, I converges. ▮

EXAMPLE 3. Consider $I = \int_0^\infty x^{100}e^{-0.01x}dx$. It might appear that this integral diverges because of the dramatic growth of the x^{100}, in spite of the $e^{-0.01x}$ decay. Let us see. Writing

$$\begin{aligned} x^{100}e^{-0.01x} &= \frac{x^{100}}{e^{0.01x}} = \frac{x^{100}}{1 + (0.01x) + \frac{(0.01x)^2}{2!} + \cdots} \\ &< \frac{x^{100}}{\frac{(0.01x)^{102}}{102!}} = \frac{(102!)10^{200}}{x^2}, \end{aligned} \tag{8}$$

we see, by comparison with the p-integral, with $p = 2$, that I converges. ∎

EXAMPLE 4. Observe that

$$I = \int_3^\infty \frac{dx}{x \ln x} = \int_{x=3}^{x=\infty} \frac{d(\ln x)}{\ln x} = \lim_{X\to\infty} \ln(\ln x)\big|_3^X = \infty, \tag{9}$$

so I is divergent. This example illustrates just how weakly $\ln x \to \infty$ as $x \to \infty$, for the integral of $1/x$ is borderline divergent ($p = 1$), and the $\ln x$ in the denominator does not even provide enough help, as $x \to \infty$, to produce borderline convergence! ∎

So much for the case where the upper limit is ∞. The other case that we consider is that in which the integrand "blows up" (i.e., tends to $+\infty$ or $-\infty$) at a finite endpoint, say the left endpoint $x = a$. If the integrand $f(x)$ blows up as $x \to a$, then in the same spirit as (3) we define

$$\boxed{I = \int_a^b f(x)\,dx \equiv \lim_{\epsilon\to 0} \int_{a+\epsilon}^b f(x)\,dx,} \tag{10}$$

where $\epsilon \to 0$ through positive values.

We first consider the so-called **vertical p-integral**

$$I = \int_0^b \frac{1}{x^p}\,dx. \qquad (b > 0) \tag{11}$$

According to (10),

$$I = \int_0^b \frac{1}{x^p}\,dx = \lim_{\epsilon\to 0} \int_\epsilon^b \frac{1}{x^p}\,dx = \begin{cases} \lim_{\epsilon\to 0} \frac{1}{1-p}\, x^{1-p}\big|_\epsilon^b & (p \neq 1) \\ \lim_{\epsilon\to 0} \ln x\big|_\epsilon^b. & (p = 1) \end{cases} \tag{12}$$

Now, $\lim_{\epsilon\to 0} \ln \epsilon$ is infinite ($-\infty$) and hence does not exist and, similarly for $\lim_{\epsilon\to 0} \epsilon^{1-p}$ if $p > 1$, whereas the latter limit does exist if $p < 1$. Thus,

THEOREM 4.5.4 *Vertical p-Integral*
The vertical p-integral, (11), converges if $p < 1$ and diverges if $p > 1$.

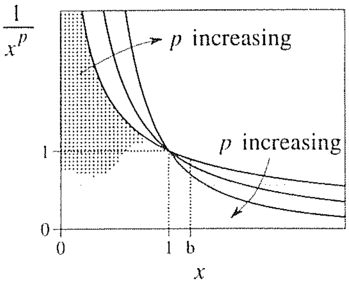


Figure 2. The effect, on $1/x^p$, of varying p.

Recall that as p is increased, the *horizontal* sliver of area (shaded in Fig. 1) is squeezed thinner and thinner. For $p > 1$ it is thin enough to have finite area. However, the effect near $x = 0$ is the opposite: increasing p causes the singularity at $x = 0$ to become stronger, and the *vertical* column of area (shaded in Fig. 2) to become thicker. Thus, to squeeze the vertical column thin enough for it to have finite area, we need to make p small enough; namely, we need $p < 1$.

The motivation behind the terms "horizontal" and "vertical" p-integrals should now be apparent; the former involves the horizontal sliver shown (shaded) in Fig. 1, and the vertical p-integral involves the vertical sliver shown (shaded) in Fig. 2.

Next, we add the following comparison test:

THEOREM 4.5.5 *Comparison Test*
Let $I = \int_0^b f(x)\,dx$, where $0 < b < \infty$. If $f(x) \sim K/x^p$ as $x \to 0$ for some constants K and p, and $f(x)$ is continuous on $0 < x \le b$, then I converges if $p < 1$ and diverges if $p \ge 1$.

EXAMPLE 5. Test the integral $\int_0^4 \left(\sin 2x/x^{3/2}\right) dx$ for convergence/divergence. Evidently, the integrand blows up as $x \to 0$ and needs to be examined there more closely. Recalling the Taylor series $\sin 2x = (2x) - (2x)^3/3! + (2x)^5/5! - \cdots$, we see that $\sin 2x \sim 2x$ as $x \to 0$ [as can be verified, if you wish, by applying l'Hôpital's rule to show that $\sin(2x)/2x \to 1$ as $x \to 0$], so

$$\frac{\sin 2x}{x^{3/2}} \sim \frac{2x}{x^{3/2}} = \frac{2}{x^{1/2}}. \tag{13}$$

Thus, according to Theorem 4.5.5, with $p = 1/2$, the integral is convergent. ▮

Example 5 concludes our introduction to singular integrals, and we are now prepared to study the gamma function.

4.5.2. Gamma function. The integral

$$\boxed{\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \qquad (x > 0)} \tag{14}$$

is nonelementary; that is, it cannot be evaluated in closed form in terms of the so-called elementary functions. Since it arises frequently, it has been given a name, the **gamma function**, and has been studied extensively.

Observe that the integral is singular for two reasons: first, the upper limit is ∞ and, second, the integrand blows up as $t \to 0$ if the exponent $x - 1$ is negative. To determine its convergence or divergence, we can separate the two singularities by breaking the integral into the sum of an integral from $t = 0$ to $t = \tau$, say, for any $\tau > 0$, plus another from τ to ∞.* In the first, we have $t^{x-1}e^{-t} \sim t^{x-1} = 1/t^{1-x}$

*That is, if $f(t)$ is unbounded as $t \to 0$, then the integral

$$\int_0^\infty f(t)\,dt \equiv \lim_{\substack{\epsilon \to 0\\ T \to \infty}} \int_\epsilon^T f(t)\,dt = \lim_{\substack{\epsilon \to 0\\ T \to \infty}} \left[\int_\epsilon^\tau f(t)\,dt + \int_\tau^T f(t)\,dt\right]$$

$$= \lim_{\epsilon \to 0} \int_\epsilon^\tau f(t)\,dt + \lim_{T \to \infty} \int_\tau^T f(t)\,dt = \int_0^\tau f(t)\,dt + \int_\tau^\infty f(t)\,dt$$

exists if and only if each of the last two integrals exist.

as $t \to 0$, and by Theorem 4.5.5 we see that we have convergence if $1-x < 1$ (i.e., $x > 0$), and divergence if $x \leq 0$. In the part from 0 to ∞, we have convergence no matter how large x is, due to the e^{-t}. Thus, the integral in (14) is convergent only if $x > 0$; hence the parenthetic stipulation in (14).

An important property of the gamma function can be derived from the definition (14) by integration by parts. With "u"$= t^{x-1}$ and "dv"$= e^{-t}dt$,

$$\Gamma(x) = -t^{x-1}e^{-t}\Big|_0^\infty + (x-1)\int_0^\infty t^{x-2}e^{-t}\,dt. \tag{15}$$

The integral in (15) converges [and is $\Gamma(x-1)$] only if $x > 1$ (rather than $x > 0$, because the exponent on t is now $x-2$), in which case the boundary term vanishes. Thus, (15) becomes

$$\boxed{\Gamma(x) = (x-1)\Gamma(x-1). \qquad (x > 1)} \tag{16}$$

The latter is a *recursion formula* because it gives Γ at one point in terms of Γ at another point. In fact, if we compute $\Gamma(x)$, by numerical integration, over a unit interval such as $0 < x \leq 1$, then (16) enables us to compute $\Gamma(x)$ for all $x > 1$. For example,

$$\Gamma(3.2) = 2.2\,\Gamma(2.2) = (2.2)(1.2)\Gamma(1.2) = (2.2)(1.2)(0.2)\Gamma(0.2), \tag{17}$$

and one can find $\Gamma(0.2)$ in a table. (Actually, tabulations are normally given over the interval $1 < x \leq 2$ because accurate integration is difficult if x is close to 0. In fact, tables are no longer essential since the gamma function is available within most computer libraries.) Note, in particular, that if n is a positive integer, then

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n) = n(n-1)\Gamma(n-1)\\ &= \cdots = n(n-1)(n-2)\cdots(1)\Gamma(1),\end{aligned} \tag{18}$$

and since

$$\Gamma(1) = \int_0^\infty e^{-t}\,dt = 1, \tag{19}$$

(18) becomes

$$\boxed{\Gamma(n+1) = n!.} \tag{20}$$

Thus, the gamma function can be evaluated analytically at positive integer values of its argument. Another x at which the integration can be carried out is $x = 1/2$, and the result is

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{21}$$

Derivation of (21) is interesting and is left for the exercises.

Recall that (14) defines $\Gamma(x)$ only for $x > 0$; for $x < 0$ the integral diverges and (14) is meaningless. What is a reasonable way to extend the definition of $\Gamma(x)$

to negative x? Recall that if we know $\Gamma(x)$, then we can compute $\Gamma(x+1)$ from the recursion formula $\Gamma(x+1) = x\Gamma(x)$. Instead of using this formula to step to the right [for instance, recall that knowing $\Gamma(0.2)$ we were able, in (17), to compute $\Gamma(3.2)$], we can turn it around and use it to step to the left. Thus, let us *define*

$$\Gamma(x) \equiv \frac{\Gamma(x+1)}{x} \qquad \text{for } x < 0. \tag{22}$$

For example,

$$\Gamma(-2.6) = \frac{\Gamma(-1.6)}{-2.6} = \frac{\Gamma(-0.6)}{(-2.6)(-1.6)} = \frac{\Gamma(0.4)}{(-2.6)(-1.6)(-0.6)}, \tag{23}$$

where $\Gamma(0.4)$ is known because its argument is positive. The resulting graph of $\Gamma(x)$ is shown in Fig. 3.

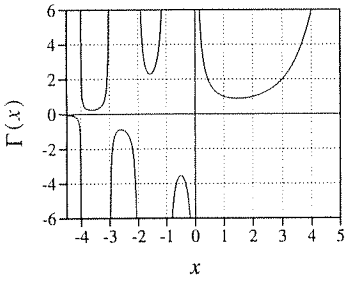


Figure 3. Gamma function, $\Gamma(x)$.

In summary then, $\Gamma(x)$ is defined for all $x \neq 0, -1, -2, \ldots$ by the integral (14) together with the leftward-stepping recursion formula (16). The singularity of $\Gamma(x)$ at $x = 0$ propagates to $x = -1, -2, \ldots$ by virtue of that formula. Especially notable is the fact that $\Gamma(x) = (x-1)!$ at $x = 1, 2, 3, \ldots$, and for this reason $\Gamma(x)$ is often referred to as the **generalized factorial function**.

A great many integrals are not themselves gamma function integrals but can be evaluated by making suitable changes of variables so as to reduce them to gamma function integrals.

EXAMPLE 6. Evaluate $I = \int_0^\infty t^{2/3} e^{-\sqrt{t}}\, dt$. Setting $\sqrt{t} = u$, we obtain

$$I = \int_0^\infty \left(u^2\right)^{2/3} e^{-u} 2u\, du = 2\int_0^\infty u^{7/3} e^{-u}\, du = 2\Gamma\left(\frac{10}{3}\right), \tag{24}$$

where $\Gamma(10/3)$ can be obtained from tables or a computer. ▮

4.5.3. Order of magnitude. In some of the foregoing examples it was important to assess the relative magnitudes of two given functions. In Example 3, for instance, the x^{100} grows as $x \to \infty$ while the $e^{-0.01x}$ decays. Which one "wins," and by what margin determines whether the integral converges or diverges.

Of particular interest are the relative growth and decay of the exponential, algebraic, and logarithmic functions as $x \to \infty$ and $x \to 0$, and we state the following elementary results as a theorem, both for emphasis and for reference.

THEOREM 4.5.6 *Relative Growth and Decay*
For any choice of positive real numbers α and β,

$$x^\alpha e^{-\beta x} \to 0 \quad \text{as } x \to \infty, \tag{25a}$$

$$(\ln x)/x^\alpha \to 0 \quad \text{as } x \to \infty, \tag{25b}$$

$$x^\alpha \ln x \to 0 \quad \text{as } x \to 0. \tag{25c}$$

Proof of (25a) can proceed as a generalization of (8) or by using l'Hôpital's rule, and (25b,c) can be proved by l'Hôpital's rule. To prove (25c), for example, observe that $x^\alpha \ln x \to (0)(-\infty)$, which result is indeterminate. To use l'Hôpital's rule we need to have $0/0$ or ∞/∞. Thus, express $x^\alpha \ln x$ as $(\ln x)/x^{-\alpha}$, which tends to $-\infty/\infty$ as $x \to 0$. Then l'Hôpital's rule gives

$$\lim_{x\to 0} \frac{\ln x}{x^{-\alpha}} = \lim_{x\to 0} \frac{x^{-1}}{-\alpha x^{-\alpha-1}} = \lim_{x\to 0} \left(-\frac{x^\alpha}{\alpha}\right) = 0.$$

We say that x^α exhibits *algebraic growth* as $x \to \infty$, and we see from (25a) that algebraic growth x^α is no match for exponential decay $e^{-\beta x}$, no matter how large α is and no matter how small β is! Of course, it follows from (25a) that $x^{-\alpha}e^{\beta x} \to \infty$ as $x \to \infty$: *algebraic decay* is no match for exponential growth. Just as exponential growth is extremely strong, logarithmic growth is extremely weak for (25b) shows that x^α dominates $\ln x$ as $x \to \infty$, no matter how small α is. Similarly as $x \to 0$: $x^{-\alpha} \to \infty$ and $\ln x \to -\infty$ (recall that $\ln x$ is zero at $x = 1$, increases without bound as $x \to \infty$, and decreases without bound as $x \to 0$; sketch it), and (25c), rewritten as $(\ln x)/x^{-\alpha} \to 0$, shows that $x^{-\alpha} \to \infty$ faster than $\ln x \to -\infty$, no matter how small α is.

Crudely then, we can think of $\ln x$ as being of the order of x to an infinitesimal positive power as $x \to \infty$, and of the order of x to an infinitesimal negative power as $x \to 0$. In contrast, one can think, crudely, of e^x as being of the order of x to an arbitrarily large positive power as $x \to \infty$, and e^{-x} as being of the order of x to an arbitrarily large negative power as $x \to \infty$.

When considering the relative strength of functions as x tends to some value x_0, constant scale factors are of no consequence no matter how large or small they may be. For instance, $(87 \ln x)/x^{0.01} \to 0$ as $x \to \infty$ just as $(\ln x)/x^{0.01}$ does. Thus, in place of the asymptotic notation $f(x) \sim g(x)$ as $x \to x_0$, which means that $f(x)/g(x) \to 1$ as $x \to x_0$, we will sometimes use the "**big oh**" notation

$$\boxed{f(x) = O(g(x)) \quad \text{as } x \to x_0} \tag{26a}$$

to mean that*

$$f(x) \sim Cg(x) \quad \text{as } x \to x_0, \tag{26b}$$

for some finite nonzero constant C. For instance, whereas

$$f(x) = \frac{\sqrt{1+\sqrt{3}}}{\Gamma(1+\sqrt{5})}\, x^{-1/2} + 673 \ln x \;\sim\; \frac{\sqrt{1+\sqrt{3}}}{\Gamma(1+\sqrt{5})}\, x^{-1/2} \tag{27}$$

as $x \to 0$, it is simpler to write $f(x) = O(x^{-1/2})$ as $x \to 0$. That is, the scale factor $C = \sqrt{1+\sqrt{3}}/\Gamma(1+\sqrt{5})$ can be omitted insofar as the order of magnitude

*Actually, the notation (26a) means that $f(x)/g(x)$ is bounded as $x \to x_0$. Though our usage is more restricted, it is consistent with the definition just given, for if (26b) holds, then surely $f(x)/g(x)$ is bounded as $x \to x_0$. Though more restricted, our definition (26b) of (26a) is sufficient for our purposes and is easier to understand and use.

of f is concerned. In words, we say that f is big oh of $x^{-1/2}$ as $x \to 0$. Of course, x_0 can be any point in (26); often, but not always, x_0 is 0 or ∞.

As one more illustration of the big oh notation, observe from the Taylor series

$$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots \tag{28}$$

that each of the following is true:

$$\sin x = O(x), \tag{29a}$$

$$\sin x = x + O(x^3), \tag{29b}$$

$$\sin x = x - \frac{x^3}{6} + O(x^5), \tag{29c}$$

and so on, as $x \to 0$. For instance, l'Hôpital's rule shows that $(\sin x)/x \to 1$ as $x \to 0$, so $\sin x \sim x$; hence (26b) holds, with $C = 1$, so (29a) is correct. Similarly, l'Hôpital's rule shows that $(\sin x - x)/x^3 \to -1/6$, so $\sin x - x \sim -x^3/6$; hence $\sin x - x = O(x^3)$ or $\sin x = x + O(x^3)$ so (29b) is correct.

The big oh notation is especially useful in working with series. For instance, (29b) states that if we retain only the leading term of the Taylor series (28), then the error thereby incurred is of order $O(x^3)$. Put differently, the portion omitted, $-\frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \cdots$, is simply $O(x^3)$.

Closure. In Section 4.5.1, we define singular integrals as integrals in which something is infinite: one or both integration limits and/or the integrand. We make such integrals meaningful by defining them as limits of regular integrals. Just as the convergence and divergence of infinite series is a long story, so is the convergence and divergence of singular integrals, but our aim here is to consider only types that will arise in this text. Though the convergence of singular integrals of the type $\int_0^\infty f(x)\,dx$ and of infinite series $\sum_1^\infty a_n$ bear a strong resemblance (e.g., the p-series and horizontal p-integral both converge for $p > 1$ and diverge for $p \leq 1$), one should by no means expect all results about infinite series to merely carry over. For instance, for series convergence it is necessary (but not sufficient) that $a_n \to 0$ as $n \to \infty$, but it is *not* necessary for the convergence of $\int_0^\infty f(x)\,dx$ that $f(x) \to 0$ as $x \to \infty$. For instance, we state without proof that $\int_0^\infty \sin(x^2)\,dx$ converges, even though $\sin(x^2)$ does not tend to zero as $x \to \infty$.

In Section 4.5.2, we introduce a specific and useful singular integral, the gamma function, and obtain its recursion formula and some of its values. The exercises indicate some of its many applications.

In the final section, 4.5.3, our aim is to clarify the relative orders of magnitude of exponential, algebraic, and logarithmic functions. It is important for you to be familiar with the results listed in Theorem 4.5.6, just as you are familiar with the relative weights of cannonballs and feathers. We also introduce a simple big oh notation which is especially useful in Chapter 6 on the numerical integration of differential equations.

Computer software. Many integrals can be evaluated by symbolic computer software. With *Maple*, for instance, the relevant command is **int**. To evaluate the integral I in Example 6, for instance, enter

$$\text{int}(t\hat{}(2/3) * \exp(-t\hat{}(1/2)),\ t = 0..\text{infinity});$$

and return. The result is

$$\frac{112}{81}\frac{\pi\sqrt{3}}{\Gamma(2/3)}$$

which looks different from the result obtained in Example 6 but is actually equivalent to that result (Exercise 12). To evaluate the latter, enter

evalf(");

and return. The result is 5.556316963.

EXERCISES 4.5

1. If $\alpha > 0$ and $\beta > 0$ show that, no matter how large α is and no matter how small β is,

(a) $x^\alpha e^{-\beta x} \to 0 \quad \text{as } x \to \infty$
(b) $x^{-\alpha} e^{\beta x} \to \infty \quad \text{as } x \to \infty$

2. If $\alpha > 0$, show that no matter how small α is,

(a) $(\ln x)/x^\alpha \to 0 \quad \text{as } x \to \infty$
(b) $x^\alpha / \ln x \to 0 \quad \text{as } x \to 0$

3. Show whether the given integral converges or diverges. As usual, be sure to explain your reasoning.

(a) $\displaystyle\int_0^\infty \frac{dx}{x^4+2}$
(b) $\displaystyle\int_0^\infty \frac{x^3\,dx}{x^5+2}$
(c) $\displaystyle\int_0^\infty \frac{x^3\,dx}{x^4+2}$
(d) $\displaystyle\int_0^\infty \frac{x^{3.2}\,dx}{x^4+100}$
(e) $\displaystyle\int_0^\infty \frac{dx}{x^2}$
(f) $\displaystyle\int_0^\infty \frac{dx}{\sqrt{x}}$
(g) $\displaystyle\int_4^\infty \frac{\sin^2 x\,dx}{\sqrt{x}(x-1)}$
(h) $\displaystyle\int_5^\infty \frac{\cos x\,dx}{x(x-1)}$
(i) $\displaystyle\int_1^2 \frac{e^x\,dx}{\sqrt{x-1}}$ HINT: Let $\xi = x - 1$.
(j) $\displaystyle\int_0^1 \frac{dx}{x^2\cos x}$

4. Show whether the given integral converges or diverges.

(a) $\displaystyle\int_1^\infty \frac{\ln x\,dx}{x^2}$ HINT: Show that $\ln x < x^{1/4}$ for all sufficiently large x, by showing that $(\ln x)/x^{1/4} \to 0$ as $x \to \infty$.
(b) $\displaystyle\int_0^2 \frac{\ln x\,dx}{\sqrt{x}}$ HINT: Make the change of variables $x = 1/\xi$ and use the hint given in part (a).

5. For what p's, if any, does $\displaystyle\int_0^\infty dx/x^p$ converge? Explain.

6. Enter the indicated change of variables in the given singular integral, and state whether the resulting integral is singular or not.

(a) $\displaystyle\int_2^\infty \frac{dx}{x^4+2}, \quad x = \frac{1}{\xi}$
(b) $\displaystyle\int_0^3 \frac{dx}{\sqrt{x}}, \quad x = \frac{1}{\xi}$
(c) $\displaystyle\int_2^\infty \frac{\cos x\,dx}{\sqrt{x}}, \quad x = \xi^2$

(d) $\displaystyle\int_2^\infty \frac{\cos x\,dx}{\sqrt{x}}, \quad x = \xi^4$

7. For what range of α's (such as $0 < \alpha < 2$, $\alpha > 4$, no α's, etc.) does the given integral converge? Explain.

(a) $\displaystyle\int_0^\infty \frac{dx}{x^\alpha + 3}$ (b) $\displaystyle\int_0^4 \frac{x^\alpha\,dx}{x^4+1}$

(c) $\displaystyle\int_0^\infty \frac{x^\alpha\,dx}{x+1}$ (d) $\displaystyle\int_0^2 x^\alpha \sin x\,dx$

(e) $\displaystyle\int_1^2 (x^2-1)^\alpha\,dx$ (f) $\displaystyle\int_1^\infty \frac{x^\alpha\,dx}{\sqrt{x^2-1}}$

8. Evaluate, using a suitable recursion formula and the known value $\Gamma(1/2) = \sqrt{\pi}$. Repeat the evaluation using computer software.

(a) $\Gamma(3.5)$ (b) $\Gamma(-3.5)$ (c) $\Gamma(6.5)$ (d) $\Gamma(-0.5)$

9. Derive (21), that $\Gamma(1/2) = \sqrt{\pi}$. HINT: Show that

$$\Gamma(1/2) = 2\int_0^\infty e^{-u^2}du,$$

so that

$$\begin{aligned}[\Gamma(1/2)]^2 &= 4\int_0^\infty e^{-u^2}du \int_0^\infty e^{-v^2}dv \\ &= 4\int_0^\infty\int_0^\infty e^{-(u^2+v^2)}du\,dv.\end{aligned}$$

Regarding the latter as a double integral in a Cartesian u,v plane, change from u,v to polar coordinates r,θ. The resulting double integral should be easier to evaluate.

10. Show by suitable change of variables that

(a) $\displaystyle\int_0^\infty e^{-x^p}dx = \frac{\Gamma(1/p)}{p} \quad (p > 0)$

(b) $\displaystyle\int_0^1 x^m (\ln x)^n\,dx = (-1)^n \frac{n!}{(m+1)^{n+1}}$
(m, n nonnegative integers)

(c) $\displaystyle\int_0^\infty x^2 e^{-x^2}dx = \frac{\sqrt{\pi}}{4}$

(d) $\displaystyle\int_0^\infty x e^{-\sqrt{x}}dx = 12$

(e) $\displaystyle\int_0^\infty (x-1)^2 e^{-x^3}dx = \frac{1}{3}\left[1 - 2\Gamma\left(\frac{2}{3}\right) + \Gamma\left(\frac{1}{3}\right)\right]$

11. Evaluate as many of the integrals in Exercise 10 as possible using computer software.

12. In Example 6 we obtained the value $2\Gamma(10/3)$. Using *Maple*, instead, show that the result is $(112/81)\pi\sqrt{3}/\Gamma(2/3)$. Then, use any formulas given in this section or in these exercises to show that the two results are equivalent.

13. Deduce, from the formula given in Exercise 10(a), that $\Gamma(x) \sim 1/x$ as x tends to zero through positive values.

14. (*Beta function*) Derive the result

$$\boxed{B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},} \qquad (14.1)$$

for $p > 0$, $q > 0$; $B(p,q)$ is known as the **beta function**. HINT: Putting $x = u^2$ in

$$\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx, \qquad (14.2)$$

show that

$$\Gamma(q)\Gamma(p) = 4\int_0^\infty u^{2p-1}e^{-u^2}du \int_0^\infty v^{2q-1}e^{-v^2}dv. \quad (14.3)$$

Regarding the latter as a double integral in a Cartesian u,v plane, change from u,v to polar coordinates r,θ. Making one more change of variables in each integral, the r integral gives $\Gamma(p+q)$ and the θ integral gives $B(p,q)$.

15. Derive, from (14.1) above, the alternative forms:

(a) $$B(p,q) = \int_0^\infty \frac{t^{p-1}}{(1+t)^{p+q}}dt \qquad (15.1)$$
$(p > 0,\ q > 0)$

HINT: Set $x = t/(1+t)$ in (14.1).

(b) $$B(p,q) = 2\int_0^{\pi/2} \cos^{2q-1}\theta\,\sin^{2p-1}\theta\,d\theta \qquad (15.2)$$
$(p > 0,\ q > 0)$

16. Using any results from the preceding two exercises, show that

(a) $$\int_0^{\pi/2} \cos^q\theta\,\sin^p\theta\,d\theta = \frac{1}{2}B\left(\frac{p+1}{2}, \frac{q+1}{2}\right) \qquad (16.1)$$
$(p > -1,\ q > -1)$

(b) $$\int_0^{\pi/2} \tan^p \theta \, d\theta = \int_0^{\pi/2} \cot^p \theta \, d\theta = \frac{1}{2} B\left(\frac{1+p}{2}, \frac{1-p}{2}\right) = \frac{\pi}{2\cos\frac{p\pi}{2}} \tag{16.2}$$

for $-1 < p < 1$. HINT: You may use (17.1), below.

(c) $$\int_0^{\infty} \frac{x^a \, dx}{(1+x^b)^c} = \frac{1}{b} B\left(\frac{a+1}{b}, \frac{cb-a-1}{b}\right) = \frac{1}{b} \frac{\Gamma\left(\frac{a+1}{b}\right)\Gamma\left(\frac{cb-a-1}{2}\right)}{\Gamma(c)} \tag{16.3}$$

for $a > -1$, $b > 0$, $bc - a > 1$.

17. It can be shown, from the residue theorem of the complex integral calculus, that

$$\int_0^{\infty} \frac{x^{\alpha-1}}{1+x} \, dx = \frac{\pi}{\sin \alpha\pi} \quad (0 < \alpha < 1) \tag{17.1}$$

Using this result, (15.1), and (14.1), show that

$$\Gamma(\alpha)\Gamma(1-\alpha) = \frac{\pi}{\sin \alpha\pi}. \quad (0 < \alpha < 1) \tag{17.2}$$

18. (*Period of oscillation of pendulum*) Conservation of energy dictates that the angular displacement $\theta(t)$ of a pendulum, of length l and mass m, satisfies the differential equation

$$\frac{1}{2} m \left(l\dot{\theta}\right)^2 + mg\,(l - l\cos\theta) = mg\,(l - l\cos\theta_0). \tag{18.1}$$

(a) From (18.1), show that

$$\sqrt{\frac{l}{2g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos\theta - \cos\theta_0}} = \int_0^{T/4} dt, \tag{18.2}$$

where T is the period and θ_0 is the maximum swing. We expect T to depend on θ_0, so we denote it as $T(\theta_0)$. For the case $\theta_0 = \pi/2$, show that

$$T(\pi/2) = \sqrt{\frac{2\pi l}{g}\frac{\Gamma(1/4)}{\Gamma(3/4)}}.$$

NOTE: You may use results from the preceding exercises.
(b) At first glance, it appears from (18.2) that $T(\theta_0) \to 0$ as $\theta_0 \to 0$. Is that assessment correct? Explain.

19. Let $F(x) = 4/(1+x^2) = 4 - 4x^2 + 4x^4 - \cdots$, $G(x) = \dfrac{2+3x}{5-4x}$, $H(x) = \dfrac{7x^3 - x + 1}{x^2+4}$, $I(x) = \dfrac{2x - 3\ln x}{1+x^2}$, $J(x) = 2e^{-x} + 3x$, and $K(x) = \sin 3x$. Verify the truth of each:

(a) $F(x) = O(1)$ as $x \to 0$
(b) $F(x) = 4 + O(x^2)$ as $x \to 0$
(c) $F(x) = 4 - 4x^2 + O(x^4)$ as $x \to 0$
(d) $G(x) = O(1)$ as $x \to \infty$
(e) $H(x) = O(x)$ as $x \to 0$
(f) $H(x) = O(x)$ as $x \to \infty$
(g) $H(x) = O(1)$ as $x \to 0$
(h) $I(x) = O(x^{-1})$ as $x \to \infty$
(i) $I(x) = O(\ln x)$ as $x \to 0$
(j) $J(x) = O(x)$ as $x \to \infty$
(k) $J(x) = O(1)$ as $x \to 0$
(l) $K(x) = O(1)$ as $x \to 0$

4.6 Bessel Functions

The differential equation

$$\boxed{x^2 y'' + xy' + \left(x^2 - \nu^2\right) y = 0,} \tag{1}$$

where ν is a nonnegative real number, is known as **Bessel's equation of order** ν. The equation was studied by *Friedrich Wilhelm Bessel* (1784–1846), director of the astronomical observatory at Königsberg, in connection with his work on planetary motion. Outside of planetary motion, the equation appears prominently in a wide range of applications such as steady and unsteady diffusion in cylindrical regions, and one-dimensional wave propagation and diffusion in variable cross-section media, and it is one of the most important differential equations in mathematical physics. Dividing through by the leading coefficient x^2, we see from $p(x) = 1/x$ and $q(x) = \left(x^2 - \nu^2\right)/x^2$ that there is one singular point, $x = 0$, and that it is a regular singular point because $xp(x) = 1$ and $x^2q(x) = x^2 - \nu^2$ are analytic at $x = 0$.

4.6.1. $\nu \neq$ integer. Consider the case where the parameter ν is not an integer. Seeking a Frobenius solution about the regular singular point $x = 0$,

$$y(x) = \sum_{k=0}^{\infty} a_k x^{k+r}, \qquad (a_0 \neq 0) \tag{2}$$

gives (Exercise 1)

$$\sum_{k=0}^{\infty} \left\{ \left[(k+r)^2 - \nu^2\right] a_k + a_{k-2} \right\} x^{k+r} = 0, \tag{3}$$

where $a_0 \neq 0$ and $a_{-2} = a_{-1} \equiv 0$. Equating to zero the coefficient of each power of x in (3) gives

$$k = 0: \quad (r^2 - \nu^2)a_0 = 0, \tag{4a}$$
$$k = 1: \quad \left[(r+1)^2 - \nu^2\right] a_1 = 0, \tag{4b}$$
$$k \geq 2: \quad \left[(r+k)^2 - \nu^2\right] a_k + a_{k-2} = 0. \tag{4c}$$

Since $a_0 \neq 0$, (4a) gives the indicial equation $r^2 - \nu^2 = 0$, with the distinct roots $r = \pm\nu$. First, let $r = +\nu$. Then (4b) gives $a_1 = 0$ and (4c) gives the recursion relation

$$a_k = -\frac{1}{k(k+2\nu)} a_{k-2}. \qquad (k \geq 2) \tag{5}$$

From (5), together with the fact that $a_1 = 0$, it follows that $a_1 = a_3 = a_5 = \cdots = 0$ and that

$$a_{2k} = \frac{(-1)^k}{2^{2k} k!\,(\nu+k)(\nu+k-1)\cdots(\nu+1)} a_0. \tag{6}$$

If ν were an integer, then the ever-growing product $(\nu+k)(\nu+k-1)\cdots(\nu+1)$ could be simplified into closed form as $\nu!/(\nu+k)!$. But since ν is not an integer, we seek to accomplish such simplification by means of the generalized factorial function, the gamma function, which is studied in Section 4.5. Specifically, if we recall the gamma function recursion formula $\Gamma(x) = (x-1)\Gamma(x-1)$, then $\Gamma(\nu+k+1) = (\nu+k)\Gamma(\nu+k) = (\nu+k)(\nu+k-1)\Gamma(\nu+k-1) = \cdots = (\nu+k)(\nu+k-1)\cdots(\nu+1)\Gamma(\nu+1)$, which gives $(\nu+k)(\nu+k-1)\cdots(\nu+1) = \Gamma(\nu+k+1)/\Gamma(\nu+1)$. With this replacement, (6) becomes

$$a_{2k} = \frac{(-1)^k \Gamma(\nu+1)}{2^{2k} k!\,\Gamma(\nu+k+1)} a_0, \tag{7}$$

so we have the solution

$$y(x) = a_0 2^\nu \Gamma(\nu+1) \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k+\nu}. \tag{8}$$

Dropping the $a_0 2^\nu \Gamma(\nu+1)$ scale factor, we call the resulting solution the **Bessel function of the first kind, of order** ν:

$$\boxed{J_\nu(x) = \left(\frac{x}{2}\right)^{\nu} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(\nu+k+1)} \left(\frac{x}{2}\right)^{2k}.} \tag{9}$$

To obtain a second linearly independent solution, we turn to the other indicial root, $r = -\nu$. There is no need to retrace all of our steps; all we need to do is to change ν to $-\nu$ everywhere on the right side of (9). Denoting the result as $J_{-\nu}(x)$, the **Bessel function of the first kind, of order** $-\nu$, we have

$$\boxed{J_{-\nu}(x) = \left(\frac{x}{2}\right)^{-\nu} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,\Gamma(k-\nu+1)} \left(\frac{x}{2}\right)^{2k}.} \tag{10}$$

Both series, (9) and (10), converge for all x, as follows from Theorem 4.3.1 with $R_1 = R_2 = \infty$ or from Theorem 4.2.2 and the recursion formula (5). The leading terms of the series in (9) and (10) are constants times x^ν and $x^{-\nu}$, respectively, so neither of the solutions $J_\nu(x)$ and $J_{-\nu}(x)$ is a scalar multiple of the other. Thus, they are LI, and we conclude that

$$y(x) = AJ_\nu(x) + BJ_{-\nu}(x) \tag{11}$$

is a general solution of (1).

Writing (9) and (10),

$$J_\nu(x) = x^\nu \left[\frac{1}{\Gamma(\nu+1)2^\nu} - \frac{1}{\Gamma(\nu+2)2^{\nu+2}}x^2 + \cdots\right], \tag{12}$$

$$J_{-\nu}(x) = x^{-\nu} \left[\frac{1}{\Gamma(1-\nu)2^{-\nu}} - \frac{1}{\Gamma(2-\nu)2^{-\nu+2}}x^2 + \cdots\right]. \tag{13}$$

Since the power series within the square brackets tend to $1/\Gamma(\nu+1)2^\nu$ and $1/\Gamma(1-\nu)2^{-\nu}$, respectively, as $x \to 0$, we see that $J_\nu(x) \sim [1/\Gamma(\nu+1)2^\nu]x^\nu$ and $J_{-\nu}(x) \sim [1/\Gamma(1-\nu)2^{-\nu}]x^{-\nu}$ as $x \to 0$. It is simpler and more concise to use the big oh notation introduced in Section 4.5.3, and say that $J_\nu(x) = O(x^\nu)$ and $J_{-\nu}(x) = O(x^{-\nu})$ as $x \to 0$. Thus, the $J_\nu(x)$'s tend to zero and the $J_{-\nu}(x)$'s tend to infinity as $x \to 0$. As representative, we have plotted $J_{1/2}(x)$ and $J_{-1/2}(x)$ in Fig. 1. In fact, for the half-integer values $\nu = \pm 1/2, \pm 3/2, \pm 5/2, \ldots$ the series in (9) and (10) can be shown to represent elementary functions. For instance (Exercise 5),

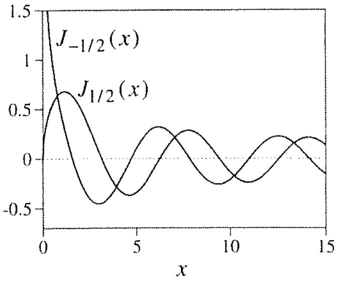


Figure 1. $J_{1/2}(x)$ and $J_{-1/2}(x)$.

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x, \qquad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x. \tag{14a,b}$$

4.6.2. ν **= integer.** If ν is a positive integer n, then $\Gamma(\nu+k+1)=\Gamma(n+k+1)=(n+k)!$ in (9), so we have from (9) the solution

$$\boxed{J_n(x)=\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,(k+n)!}\left(\frac{x}{2}\right)^{2k+n}} \tag{15}$$

of (1). For instance,

$$J_0(x)=1-\frac{x^2}{2^2}+\frac{x^4}{2^4(2!)^2}-\frac{x^6}{2^6(3!)^2}+\cdots, \tag{16a}$$

$$J_1(x)=\frac{x}{2}-\frac{x^3}{2^3 2!}+\frac{x^5}{2^5 2!\,3!}-\frac{x^7}{2^7 3!\,4!}+\cdots. \tag{16b}$$

We need to be careful with (10) because if $\nu=n$, then the $\Gamma(k-n+1)$ in (10) is, we recall from Section 4.5.2, undefined when its argument is zero or a negative integer – namely, for $k=0,1,\ldots,n-1$. One could say that $\Gamma(k-n+1)$ is infinite for those k's, so $1/\Gamma(k-n+1)$ equals zero for $k=0,1,\ldots,n-1$, and it equals $1/(k-n)!$ for $k=n,n+1,\ldots$, in which case (10) becomes

$$J_{-n}(x)=\sum_{k=n}^{\infty}\frac{(-1)^k}{k!\,(k-n)!}\left(\frac{x}{2}\right)^{2k-n}. \tag{17}$$

[The resulting equation (17) is correct, but our reasoning was not rigorous since $\Gamma(k-n+1)$ is *undefined* at $k=0,1,\ldots,n-1$, rather than "∞." A rigorous line of approach is suggested in Exercise 10.] Replacing the dummy summation index k by m according to $k-n=m$,

$$J_{-n}(x)=\sum_{m=0}^{\infty}\frac{(-1)^{m+n}}{(m+n)!\,m!}\left(\frac{x}{2}\right)^{2m+n}.$$

If $(-1)^n$ is factored out, the series that remains is the same as that given in (15), so that

$$J_{-n}(x)=(-1)^n J_n(x). \tag{18}$$

The result is that $J_n(x)$ and $J_{-n}(x)$ are linearly dependent, since (18) tells us that one is a scalar multiple of the other. Thus, whereas $J_\nu(x)$ and $J_{-\nu}(x)$ are LI and give the general solution (11) if ν is not an integer, we have only one linearly independent solution thus far for the case where $\nu=n$, namely, $y_1(x)=J_n(x)$ given by (15). To obtain a second LI solution $y_2(x)$ we rely on Theorem 4.3.1.

Let us begin with $n=0$. Then we have the case of repeated indicial roots, $r=\pm n=\pm 0$, which corresponds to case (ii) of that theorem. Accordingly, we seek $y_2(x)$ in the form

$$y_2(x)=J_0(x)\ln x+\sum_{1}^{\infty}c_k x^k. \tag{19}$$

Doing so, we can evaluate the c_k's, and we obtain

$$y_2(x) = J_0(x)\ln x + \left(\frac{x}{2}\right)^2 - \left(1+\frac{1}{2}\right)\frac{1}{(2!)^2}\left(\frac{x}{2}\right)^4 + \cdots, \tag{20}$$

which is called $\mathbf{Y}_0(x)$, the Neumann function of order zero. Thus, Theorem 4.3.1 leads us to the two LI solutions $y_1(x) = J_0(x)$ and $y_2(x) = \mathbf{Y}_0(x)$, so we can use them to form a general solution of (1). However, following Weber, it proves to be convenient and standard to use, in place of $\mathbf{Y}_0(x)$, a linear combination of $J_0(x)$ and $\mathbf{Y}_0(x)$, namely,

$$y_2(x) = \frac{2}{\pi}\left[\mathbf{Y}_0(x) + (\gamma - \ln 2)J_0(x)\right] \equiv Y_0(x), \tag{21}$$

where

$$Y_0(x) = \frac{2}{\pi}\left[\left(\ln\frac{x}{2} + \gamma\right)J_0(x) + \frac{x^2}{2^2} - \left(1+\frac{1}{2}\right)\frac{x^4}{2^4(2!)^2} + \left(1+\frac{1}{2}+\frac{1}{3}\right)\frac{x^6}{2^6(3!)^2} - \cdots\right] \tag{22}$$

is Weber's **Bessel function of the second kind, of order zero**; $\gamma = 0.5772157$ is known as **Euler's constant** and is sometimes written as C, and $Y_0(x)$ is sometimes written as $N_0(x)$. The graphs of $J_0(x)$ and $Y_0(x)$ are shown in Fig. 2. Important features are that $J_0(x)$ and $Y_0(x)$ look a bit like damped cosine and sine functions, except that $Y_0(x)$ tends to $-\infty$ as $x \to 0$. Specifically, we see from (16a) and (22) that

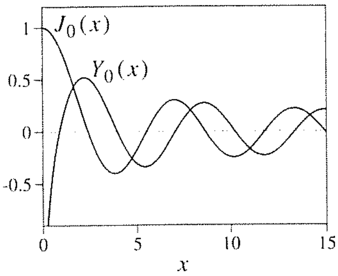


Figure 2. J_0 and Y_0.

$$J_0(x) \sim 1, \qquad Y_0(x) \sim \frac{2}{\pi}\ln x \tag{23a,b}$$

as $x \to 0$, and it can be shown (Exercise 6) that

$$J_0(x) \sim \sqrt{\frac{2}{\pi x}}\cos\left(x - \frac{\pi}{4}\right), \qquad Y_0(x) \sim \sqrt{\frac{2}{\pi x}}\sin\left(x - \frac{\pi}{4}\right) \tag{24a,b}$$

as $x \to \infty$. Indeed, we can see from (24) why the Weber Bessel function Y_0 is a nicer companion for J_0 than the Neumann Bessel function $\mathbf{Y}_0$, for

$$\mathbf{Y}_0(x) \sim \frac{1}{\sqrt{\pi x}}\left[\left(\frac{\pi}{2} - \gamma + \ln x\right)\sin x - \left(\frac{\pi}{2} + \gamma - \ln x\right)\cos x\right] \tag{24c}$$

as $x \to \infty$; surely (24b) makes a nicer companion for (24a) than does (24c).

It might appear, from Fig. 2 and (24), that the *zeros* of J_0 and Y_0 [i.e., the roots of $J_0(x) = 0$ and $Y_0(x) = 0$] are equally spaced, but they are not; they approach an equal spacing only as $x \to \infty$. For instance, the first several zeros of J_0 are 2.405, 5.520, 8.654, 11.792, 14.931. Their differences are 3.115, 3.134, 3.138, 3.139, and these are seen to rapidly approach a constant [namely, π, the spacing between the zeros of $\cos(x - \pi/4)$ in (22a)]. The zeros of the various Bessel functions turn out to be important, and they are tabulated to many significant figures.

For $n = 1, 2, \ldots$ the indicial roots $r = \pm n$ differ by an integer, which corresponds to case (iii) of Theorem 4.3.1. Using that theorem, and the ideas given above for Y_0 we obtain Weber's **Bessel function of the second kind, of order** n,

$$\boxed{\begin{aligned} Y_n(x) = \frac{2}{\pi}\Bigg[&\left(\ln\frac{x}{2}+\gamma\right)J_n(x) - \frac{1}{2}\sum_{k=0}^{n-1}\frac{(n-k-1)!}{k!}\left(\frac{x}{2}\right)^{2k-n} \\ &-\frac{1}{2}\sum_{k=0}^{\infty}\frac{(-1)^k\,[\phi(k)+\phi(k+n)]}{k!\,(k+n)!}\left(\frac{x}{2}\right)^{2k+n}\Bigg], \end{aligned}} \tag{25}$$

which formula holds for $n = 0, 1, 2, \ldots$; $\phi(0) = 0$ and $\phi(k) = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}$ for $k \geq 1$.

4.6.3. General solution of Bessel equation. Thus, we have two different general solution forms for (1), depending on whether ν is an integer or not: $y(x) = AJ_\nu(x) + BJ_{-\nu}(x)$ if ν is not an integer, and $y(x) = AJ_n(x) + BY_n(x)$ if $\nu = n = 0, 1, 2, \ldots$. It turns out that if we define

$$\boxed{Y_\nu(x) \equiv \frac{(\cos\nu\pi)\,J_\nu(x) - J_{-\nu}(x)}{\sin\nu\pi}} \tag{26}$$

for noninteger ν, then the limit of $Y_\nu(x)$ as $\nu \to n$ $(n = 0, 1, 2, \ldots)$ gives the same result as (25). Furthermore, $J_\nu(x)$ and $Y_\nu(x)$ are LI (Exercise 11) so the upshot is that we can express the general solution of (1) as

$$\boxed{y(x) = AJ_\nu(x) + BY_\nu(x)} \tag{27}$$

for *all* values of ν, with Y_ν defined by (25) and (26) for integer and noninteger values of ν, respectively. The graphs of several J_n's and Y_n's are shown in Fig. 3.

For reference, we cite the following asymptotic behavior:

$$J_n(x) \sim \frac{1}{2^n n!}x^n, \qquad (n = 0, 1, 2, \ldots) \tag{28a}$$

$$Y_n(x) \sim \begin{cases} \dfrac{2}{\pi}\ln x, & (n = 0) \\ -\dfrac{2^n(n-1)!}{\pi}\dfrac{1}{x^n}, & (n = 1, 2, \ldots) \end{cases} \tag{28b}$$

as $x \to 0$, and

$$J_n(x) \sim \sqrt{\frac{2}{\pi x}}\cos(x - \psi_n), \qquad (n = 0, 1, 2, \ldots) \tag{29a}$$

$$Y_n(x) \sim \sqrt{\frac{2}{\pi x}}\sin(x - \psi_n), \qquad (n = 0, 1, 2, \ldots) \tag{29b}$$

as $x \to \infty$, where $\psi_n = (2n+1)\pi/4$. Observe the sort of conservation expressed in (28a,b): as n increases, the Y_n's develop stronger singularities ($\ln x$, x^{-1}, $x^{-2}, \ldots$)

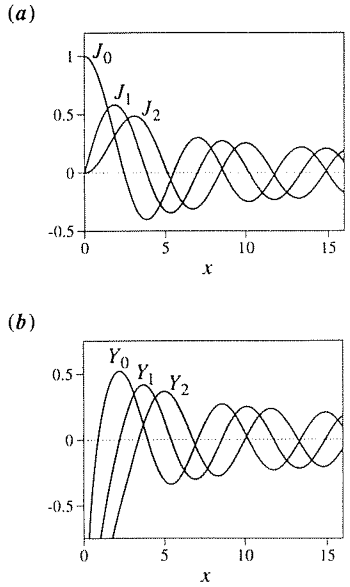


Figure 3. J_n's and Y_n's.

while the J_n's develop stronger zeros $(1, x, x^2, \ldots)$. (We say that x^5 has a stronger zero at the origin than x^3, for example, because $x^5/x^3 \to 0$ as $x \to 0$.) Finally, we call attention to the interlacing of the zeros of J_n and Y_n. All of these features can be seen in Fig. 3.

In summary, our key result is the general solution (27) with J_ν given by (9), Y_ν by (25) for integer ν and by (26) for noninteger ν, and with $J_{-\nu}$ in (26) given by (10).

4.6.4. Hankel functions. (Optional) Recall that the harmonic oscillator equation $y'' + y = 0$ has two preferred bases: $\cos x$, $\sin x$ and e^{ix}, e^{-ix}. Usually, the former is used because those functions are real valued, but sometimes the complex exponentials are more convenient. The connection between them is given by Euler's formulas:

$$\begin{aligned} e^{ix} &= \cos x + i \sin x, \\ e^{-ix} &= \cos x - i \sin x. \end{aligned}$$

Analogous to the complex basis e^{ix}, e^{-ix} for the equation $y'' + y = 0$, a complex-valued basis is defined for the Bessel equation (1):

$$H_\nu^{(1)}(x) \equiv J_\nu(x) + i\, Y_\nu(x) \tag{30a}$$
$$H_\nu^{(2)}(x) \equiv J_\nu(x) - i\, Y_\nu(x). \tag{30b}$$

These are called the **Hankel functions** of the first and second kind, respectively, of order ν. Thus, alternatively to (27), we have the general solution

$$\boxed{y(x) = AH_\nu^{(1)}(x) + BH_\nu^{(2)}(x)} \tag{31}$$

of (1).

As a result of (29a,b), the Hankel functions $H_n^{(1)}(x)$, $H_n^{(2)}(x)$ have the pure complex exponential behavior

$$H_n^{(1)}(x) \sim \sqrt{\frac{2}{\pi x}}\, e^{i(x-\psi_n)}, \tag{32a}$$
$$H_n^{(2)}(x) \sim \sqrt{\frac{2}{\pi x}}\, e^{-i(x-\psi_n)} \tag{32b}$$

as $x \to \infty$.

The Hankel functions are particularly useful in the study of wave propagation.

4.6.5. Modified Bessel equation. Besides the Bessel equation of order ν, one also encounters the **modified Bessel equation of order** ν, $x^2y''+xy'+\left(-x^2 - \nu^2\right) y = 0$, where the only difference is the minus sign in front of the second x^2 term. Let us limit our attention, for brevity, to the case where ν is an integer n, so we have

$$x^2y'' + xy' + \left(-x^2 - n^2\right) y = 0. \tag{33}$$

The change of variables $t = ix$ (or $x = -it$) converts (33) to the Bessel equation

$$t^2Y'' + tY' + \left(t^2 - n^2\right) Y = 0, \tag{34}$$

where $y(x) = y(-it) \equiv Y(t)$ and the primes on Y denote d/dt. Since a general solution to (34) is $Y(t) = AJ_n(t) + BY_n(t)$ we have, immediately, the general solution

$$y(x) = AJ_n(ix) + BY_n(ix) \tag{35}$$

of (33). From (15),

$$J_n(ix) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\,(k+n)!} \left(\frac{ix}{2}\right)^{2k+n} = i^n \sum_{k=0}^{\infty} \frac{1}{k!\,(k+n)!} \left(\frac{x}{2}\right)^{2k+n} \tag{36}$$

so we can absorb the i^n into A and be left with the real-valued solution

$$\boxed{I_n(x) \equiv i^{-n} J_n(ix) = \sum_{k=0}^{\infty} \frac{1}{k!\,(k+n)!} \left(\frac{x}{2}\right)^{2k+n},} \tag{37}$$

known as the **modified Bessel function of the first kind, and order** n. In place of $Y_n(ix)$ it is standard to introduce, as a second real-valued solution, the **modified Bessel function of the second kind, and order** n,

$$K_n(x) = \frac{\pi}{2} i^{n+1} \left[J_n(ix) + i\,Y_n(ix)\right]. \tag{38}$$

For instance,

$$I_0(x) = 1 + \frac{x^2}{2^2} + \frac{x^4}{2^4(2!)^2} + \frac{x^6}{2^6(3!)^2} + \cdots \tag{39a}$$

$$\begin{aligned} K_0(x) = &- \left(\ln \frac{x}{2} + \gamma\right) I_0(x) + (1)\frac{x^2}{2^2(1!)^2} \\ &+ \left(1 + \frac{1}{2}\right) \frac{x^4}{2^4(2!)^2} + \left(1 + \frac{1}{2} + \frac{1}{3}\right) \frac{x^6}{2^6(3!)^2} + \cdots, \end{aligned} \tag{39b}$$

and the graphs of these functions are plotted in Fig. 4.

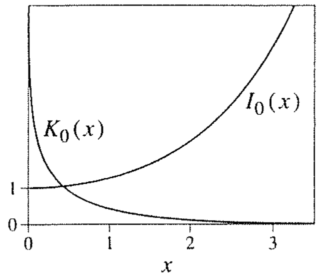


Figure 4. I_0 and K_0.

As a general solution of (33) we have

$$\boxed{y(x) = AI_n(x) + BK_n(x).} \tag{40}$$

Whereas the Bessel functions are oscillatory, the modified Bessel functions are not.

To put the various Bessel functions in perspective, observe that the relationship between the J, Y solutions of the Bessel equation and the I, K solutions of the modified Bessel equation is similar to that between the solutions $\cos x$, $\sin x$ of the harmonic oscillator equation $y'' + y = 0$ and the $\cosh x$, $\sinh x$ solutions of the "modified harmonic oscillator" equation $y'' - y = 0$. For instance, just as

$\cos(ix) = \cosh(x)$ and $\sin(ix) = i\sinh(x)$, $I_n(x)$ and $K_n(x)$ are linear combinations of $J_n(ix)$ and $Y_n(ix)$.

Finally, the asymptotic behavior of I_n and K_n is as follows:

$$I_n(x) \sim \frac{1}{n!}\left(\frac{x}{2}\right)^n, \qquad (n = 0, 1, 2, \ldots) \tag{41a}$$

$$K_n(x) \sim \begin{cases} -\ln x & (n = 0) \\ \frac{(n-1)!}{2}\left(\frac{2}{x}\right)^n, & (n = 1, 2, \ldots) \end{cases} \tag{41b}$$

as $x \to 0$, and

$$I_n(x) \sim \frac{1}{\sqrt{2\pi x}}e^x, \qquad (n = 0, 1, 2, \ldots) \tag{42a}$$

$$K_n(x) \sim \sqrt{\frac{\pi}{2x}}e^{-x}, \qquad (n = 0, 1, 2, \ldots) \tag{42b}$$

as $x \to \infty$. As n increases, the I_n's develop stronger zeros at $x = 0$ $(1, x, x^2, \ldots$ while the K_n's develop stronger singularities there $(\ln x, x^{-1}, x^{-2}, \ldots)$.

4.6.6. Equations reducible to Bessel equations. The results discussed in Sections 4.6.1–4.6.5 are all the more important because there are many equations which, although they are not Bessel equations or modified Bessel equations, can be reduced to Bessel or modified Bessel equations by changes of variables and then solved in closed form in terms of Bessel or modified Bessel functions.

EXAMPLE 1. Solve

$$xy'' + y' + \kappa^2 xy = 0. \tag{43}$$

Equation (43) is not quite a Bessel equation of order zero because of the κ^2. Let us try to absorb the κ^2 by a change of variable. Specifically, scale x as $t = \alpha x$, where α is to be determined. Then $\frac{d}{dx} = \left(\frac{d}{dt}\right)\left(\frac{dt}{dx}\right) = \alpha\frac{d}{dt}$, so $\frac{d^2}{dx^2} = \alpha^2\frac{d^2}{dt^2}$ and (43) becomes, after division by α,

$$tY'' + Y' + \frac{\kappa^2}{\alpha^2}tY = 0, \tag{44}$$

where $y(x) = y(t/\alpha) \equiv Y(t)$. Thus, we can absorb the κ^2 in (44) by choosing $\alpha = \kappa$. Then (44) is a Bessel equation of order zero with general solution $Y(t) = AJ_0(t) + BY_0(t)$ so

$$y(x) = AJ_0(t) + BY_0(t) = AJ_0(\kappa x) + BY_0(\kappa x) \tag{45}$$

is a general solution of (43). ▮

More generally, the equation

$$\boxed{\frac{d}{dx}\left(x^a\frac{dy}{dx}\right) + bx^c y = 0,} \tag{46}$$

where a, b, c are real numbers, can be transformed to a Bessel equation by transforming both independent and dependent variables. Because of the powers of x in (46), it seems promising to change variables from $x, y(x)$ to $t, u(t)$ according to the forms $t = Ax^B$, $u = x^C y$, and to try to find A, B, C so that the new equation, on $u(t)$, is a Bessel equation. It turns out that that plan works and one finds that under the change of variables

$$t = \alpha\sqrt{b}\, x^{1/\alpha} \qquad \text{and} \qquad u = x^{-\nu/\alpha} y \tag{47}$$

equation (46) becomes the Bessel equation of order ν,

$$t^2 \frac{d^2u}{dt^2} + t\frac{du}{dt} + \left(t^2 - \nu^2\right) u = 0, \tag{48}$$

if we choose

$$\alpha = \frac{2}{c-a+2} \qquad \text{and} \qquad \nu = \frac{1-a}{c-a+2}. \tag{49}$$

[The latter is meaningless if $c-a+2=0$, but in that case (46) is merely a Cauchy-Euler equation.] Thus, if Z_ν denotes any Bessel function solution of (48), then putting (47) into $u(t) = Z_\nu$ and solving for y gives the solution

$$\boxed{y(x) = x^{\nu/\alpha} Z_\nu\left(\alpha\sqrt{|b|}\, x^{1/\alpha}\right)} \tag{50}$$

of (46). If $b > 0$, then Z_ν denotes J_ν and Y_ν, and if $b < 0$, then Z_ν denotes I_ν and K_ν (though we gave formulas for I_ν and K_ν only for the case where ν is an integer).

EXAMPLE 2. Solve the equation

$$y'' + 3\sqrt{x}\, y = 0. \qquad (0 < x < \infty) \tag{51}$$

Comparing (51) with (46), we see that $a = 0$, $b = 3$, and $c = 1/2$, so $\alpha = 2/(1/2-0+2) = 4/5$ and $\nu = 1/(1/2 - 0 + 2) = 2/5$. Thus, (50) gives

$$y(x) = x^{1/2} Z_{2/5}\left(\frac{4}{5}\sqrt{3}\, x^{5/4}\right), \tag{52}$$

and

$$y(x) = \sqrt{x}\left[AJ_{2/5}\left(\frac{4}{5}\sqrt{3}\, x^{5/4}\right) + BY_{2/5}\left(\frac{4}{5}\sqrt{3}\, x^{5/4}\right)\right] \tag{53}$$

is a general solution of (51). ▮

EXAMPLE 3. Solve

$$xy'' + 3y' + y = 0, \qquad (0 < x < \infty) \tag{54}$$

or

$$y'' + \frac{3}{x}y' + \frac{1}{x}y = 0. \tag{55}$$

Writing out (46) as $x^a y'' + ax^{a-1}y' + bx^c y = 0$ or

$$y'' + \frac{a}{x}y' + bx^{c-a}y = 0, \tag{56}$$

and comparing (56) and (55) term by term gives $a = 3$, $b = 1$, and $c - a = -1$, so $c = 2$. Hence, $\alpha = 2/(2 - 3 + 2) = 2$ and $\nu = (1 - 3)/(2 - 3 + 2) = -2$, so (50) becomes

$$y(x) = x^{-2/2} Z_{-2}\left(2\sqrt{1}\,x^{1/2}\right) = x^{-1} Z_2\left(2\sqrt{x}\right) \tag{57}$$

and

$$y(x) = \frac{1}{x}\left[AJ_2\left(2\sqrt{x}\right) + BY_2\left(2\sqrt{x}\right)\right] \tag{58}$$

is a general solution of (54). ▮

NOTE: In the second equality of (57) we changed the Z_{-2} to Z_2. More generally, if the ν that we compute in (49) turns out to be negative we can always change the Z_ν in (50) to $Z_{|\nu|}$, for if ν is a negative integer $-n$, then the Z_ν in (50) gives J_{-n} and Y_{-n}; but (18) told us that J_{-n} is identical to J_n, to within a constant scale factor, and it can likewise be shown that Y_{-n} is identical to Y_n, to within a constant scale factor [namely, $Y_{-n}(x) = (-1)^n Y_n(x)$]. And if ν is negative but not an integer, then the Z_ν in (50) gives J_ν and $J_{-\nu}$, and that is equivalent to $Z_{-\nu}$ giving $J_{-\nu}$ and J_ν.

EXAMPLE 4. Solve

$$(xy')' - 5x^3 y = 0. \tag{59}$$

We see from (46) that $a = 1$, $b = -5$, $c = 3$, so $\alpha = 1/2$, $\nu = 0$ and

$$y(x) = x^0 Z_0\left(\frac{1}{2}\sqrt{|-5|}\,x^2\right),$$

so

$$y(x) = AI_0\left(\frac{\sqrt{5}}{2}x^2\right) + BK_0\left(\frac{\sqrt{5}}{2}x^2\right) \tag{60}$$

is a general solution of (59). ▮

Closure. In this section we studied the Bessel equation

$$x^2y'' + xy' + \left(x^2 - \nu^2\right)y = 0 \tag{61}$$

and the modified Bessel equation

$$x^2y'' + xy' + \left(-x^2 - \nu^2\right)y = 0. \tag{62}$$

For heuristic purposes, it is useful to keep in mind the similarity between the Bessel equation and the harmonic oscillator equation

$$y'' + y = 0, \tag{63}$$

and between the modified Bessel equation and the "modified harmonic oscillator" equation

$$y'' - y = 0. \tag{64}$$

For large x, the left side of (61) becomes

$$y'' + \frac{1}{x}y' + \left(1 - \frac{\nu^2}{x^2}\right) y \approx y'' + y \tag{65}$$

so we expect qualitative similarity between the solutions of (61) and those of (63). In fact, the solutions $J_\nu(x)$ and $Y_\nu(x)$ of (61) do tend to harmonic functions as $x \to \infty$, like the cosine and sine solutions of (63), and the y'/x term in (65) causes some damping of those harmonic functions, by a factor of $1/\sqrt{x}$. Thus, the general solution

$$y(x) = AJ_\nu(x) + BY_\nu(x) \tag{66}$$

of (61) is similar, qualitatively, to that of (63). Further, just as one can use pure complex exponential solutions of (63) according to the Euler definitions, one can introduce the Hankel functions in essentially the same way, and write the general solution of (61), alternatively, as

$$y(x) = AH_\nu^{(1)}(x) + BH_\nu^{(1)}(x). \tag{67}$$

Likewise, for the modified Bessel equation (62), the left side of which becomes

$$y'' + \frac{1}{x}y' + \left(1 - \frac{\nu^2}{x^2}\right) y \approx y'' + y \tag{68}$$

for large x, we find nonoscillatory solutions analogous to the hyperbolic cosine and sine solutions of (64).

So much for large x. As $x \to 0$, the Y_ν solutions of (61) are unbounded as are the K_ν solutions of (62).

Computer software. As a general rule of thumb, if we can derive a solution to a given differential equation by hand, we can probably obtain it using computer software. For instance, if, using *Maple*, we attempt to solve the nonconstant-coefficient differential equation (54) by the command

$$\text{dsolve}(x * \text{diff}(y(x), x, x) + 3 * \text{diff}(y(x), x) + y(x) = 0,\ y(x));$$

we do obtain the same general solution as was obtained here in Example 3.

EXERCISES 4.6

1. Putting the solution form (2) into the Bessel equation (1), derive the recursion relation (3).

2. Solve (1) for the case where $\nu = 1/2$, by the method of Frobenius. Show that your two LI solutions can be expressed in closed form as given in (14a,b).

3. Show that with $Y_0(x)$ defined by (21), the asymptotic behavior given in (24b) follows from (24a) and (24c).

4. (*Recursion formulas*) It can be shown that

$$\frac{d}{dx}\left[x^{\nu} Z_{\nu}(x)\right] = \begin{cases} x^{\nu} Z_{\nu-1}(x), & (Z = J, Y, I, H^{(1)}, H^{(2)}) \\ -x^{\nu} Z_{\nu-1}(x), & (Z = K) \end{cases} \tag{4.1}$$

$$\frac{d}{dx}\left[x^{-\nu} Z_{\nu}(x)\right] = \begin{cases} -x^{-\nu} Z_{\nu+1}(x), & (Z = J, Y, K, H^{(1)}, H^{(2)}) \\ x^{-\nu} Z_{\nu+1}(x), & (Z = I) \end{cases} \tag{4.2}$$

where the "$Z = \cdots$" means that the formula holds with Z equal to each of the itemized functions. In particular, the formula

$$\frac{d}{dx} Z_0(x) = \begin{cases} -Z_1(x), & (Z = J, Y, K, H^{(1)}, H^{(2)}) \\ Z_1(x), & (Z = I) \end{cases} \tag{4.3}$$

corresponding to (4.2) for the case $\nu = 0$, is useful in evaluating certain Bessel function integrals by integration by parts.

(a) Verify (4.1) and (4.2) for the case where Z is J (i.e., J_ν, $J_{-\nu}$ and J_n) using the series given in the text for those functions.
(b) From the formulas given above, show that

$$Z_{\nu+1}(x) = \frac{2\nu}{x} Z_{\nu}(x) - Z_{\nu-1}(x), \quad \left(Z = J, Y, H^{(1)}, H^{(2)}\right) \tag{4.4}$$

and

$$I_{\nu+1}(x) = -\frac{2\nu}{x} I_{\nu}(x) - I_{\nu-1}(x), \tag{4.5}$$

$$K_{\nu+1}(x) = \frac{2\nu}{x} K_{\nu}(x) + K_{\nu-1}(x). \tag{4.6}$$

(c) Use computer software to differentiate $x^3 J_3(x)$, $x^2 Y_2(x)$, $x^5 I_5(x)$, $x^{-2} K_2(x)$, $J_0(x)$, $Y_0(x)$, $I_0(x)$, and $K_0(x)$, and show that the results are in accord with the formulas (4.1)–(4.3).

5. (*Half-integer formulas*) (a) Putting $\nu = 1/2$ in (9) and (10), show that they give

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\sin x \tag{5.1}$$

and

$$J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\cos x. \tag{5.2}$$

(b) Derive, from (4.4), the recursion formula

$$J_{n+1/2}(x) = \frac{2n-1}{x} J_{n-1/2}(x) - J_{n-3/2}(x). \tag{5.3}$$

(c) It follows from (5.1)–(5.3) that *all* $J_{n+1/2}$'s are expressible in closed form in terms of $\sin x$, $\cos x$, and powers of x. Derive those expressions for $J_{3/2}(x)$ and $J_{-3/2}(x)$.

6. (a) (*Normal form*) By making the change of variables $y(x) = \sigma(x)v(x)$, from y to v, in $y'' + p(x)y' + q(x)y = 0$, show that the first derivative term can be eliminated by choosing σ such that $2\sigma' + p\sigma = 0$. Show that the result is the *normal form* (i.e., canonical or simplest form)

$$v'' + \left(\frac{\sigma'' + p\sigma' + q\sigma}{\sigma}\right) v = 0, \tag{6.1}$$

where

$$\sigma(x) = e^{-\frac{1}{2}\int^{x} p(t)\,dt}. \tag{6.2}$$

(b) (*Large-x behavior of Bessel functions*) For the Bessel equation (1), show that $\sigma(x) = 1/\sqrt{x}$ and that (6.1) is

$$v'' + \left(1 - \frac{\nu^2 - 1/4}{x^2}\right) v = 0. \tag{6.3}$$

NOTE: If we write $1 - (\nu^2 - 1/4)/x^2 \approx 1$ for large x, then (6.3) becomes $v'' + v \approx 0$, so we expect that $v(x) \approx A\cos(x + \phi)$ or, equivalently, $A\sin(x + \phi)$, where

A and ϕ are arbitrary constants. Thus,

$$y(x) = \sigma(x)v(x) \approx \frac{A}{\sqrt{x}}\cos(x+\phi) \quad \text{or} \quad \frac{A}{\sqrt{x}}\sin(x+\phi), \tag{6.4}$$

which forms are the same as those given by (24a) and (24b). Thus, we expect every solution of (1) to behave according to (6.4) as $x \to \infty$. Evaluating the constants A and ϕ corresponding to a particular solution, such as $J_0(x)$ or $Y_0(x)$, is complicated and will not be discussed here.

7. Recall from Example 1 that $J_n(\kappa x)$ satisfies the differential equation $x^2y'' + xy' + (\kappa^2x^2 - n^2)y = 0$ or, equivalently,

$$(xy')' + \left(\kappa^2 x - \frac{n^2}{x}\right)y = 0. \quad (n = 0, 1, 2, \dots) \tag{7.1}$$

Let the x interval be $0 < x < c$, and suppose that κ is chosen so that $J_n(\kappa c) = 0$; i.e., κc is any of the zeros of $J_n(x) = 0$. The purpose of this exercise is to derive the formula

$$\boxed{\int_0^c [J_n(\kappa x)]^2\, x\, dx = \frac{c^2}{2}[J_{n+1}(\kappa c)]^2,} \tag{7.2}$$

which will be be needed when we show how to use the Sturm-Liouville theory to expand functions on $0 < x < c$ in terms of Bessel functions. In turn, that concept will be needed later in our study of partial differential equations. To derive (7.2), we suggest the following steps.

(a) Multiplying (7.1) by $2xy'$ and integrating on x from 0 to c, obtain

$$(xy')^2\Big|_0^c + 2\int_{x=0}^{x=c}\left(\kappa^2x^2 - n^2\right)y\,dy = 0. \tag{7.3}$$

(b) Show that with $y = J_n(\kappa x)$, the $(xy')^2$ term is zero at $x = 0$ for $n = 0, 1, 2, \dots$, and that at $x = c$ it is $c^2\kappa^2[J_{n+1}(\kappa c)]^2$. HINT: It follows from (4.2) that $J_n'(x) = -J_{n+1}(x) + \frac{n}{x}J_n(x)$.

(c) Thus, show that (7.3) reduces to

$$c^2\kappa^2\left[J_{n+1}(\kappa c)\right]^2 + 2\kappa^2\int_{x=0}^{x=c} x^2y\,dy - n^2y^2\Big|_{x=0}^{x=c} = 0. \tag{7.4}$$

(d) Show that the $n^2y^2\big|_{x=0}^{x=c}$ term is zero for any $n = 0, 1, 2, \dots$, integrate the remaining integral by parts and show that the resulting boundary term is zero, and thus obtain the desired result (7.2).

8. (*Generating function for J_n*) Just as there is a "generating function" for the Legendre polynomials [see (9) in Section 4.4], there is one for the Bessel functions. Specifically, it can be shown that

$$\boxed{e^{\frac{x}{2}\left(t - \frac{1}{t}\right)} = \sum_{n=-\infty}^{\infty} J_n(x)t^n,} \tag{8.1}$$

and the left-hand side is called the **generating function** for the J_n's.

(a) We do not ask you to derive (8.1) but only to verify the $n = 0$ term. That is, expanding $e^{xt/2}$ and $e^{-x/2t}$ in Maclaurin series (one in ascending powers of t and one in powers of $1/t$) and multiplying these series together, show that the coefficient of t^0 is $J_0(x)$.

(b) Equation (8.1) is useful for deriving various properties of the J_n's. For example, taking $\partial/\partial x$ of both sides, show that

$$\frac{d}{dx}J_n(x) = \frac{1}{2}\left[J_{n-1}(x) - J_{n+1}(x)\right]. \quad (n = 1, 2, \dots) \tag{8.2}$$

(c) Similarly, taking $\partial/\partial t$ of both sides, show that

$$J_{n+1}(x) = \frac{x}{2(n+1)}\left[J_n(x) + J_{n+2}(x)\right]. \tag{8.3}$$

(d) Using computer software, differentiate $J_0(x)$ and $J_1(x)$ and show that the results agree with (8.2).

9. (*Integral representation of J_n*) Besides the generating function (preceding exercise), another source of information about the J_n's is the integral representation

$$\boxed{J_n(x) = \frac{1}{\pi}\int_0^{\pi}\cos(n\theta - x\sin\theta)d\theta.} \tag{9.1}$$

Verify (9.1) for the case $n = 0$ by using the Taylor series of $\cos t$, where $t = n\theta - x\sin\theta$ and integrating term by term. HINT: You may use any of the formulas given in the exercises to Section 4.5.

10. To derive (17) from (10), we argued that $1/\Gamma(k-n+1) = 0$ for $k = 0, 1, \dots, n-1$, on the grounds that for those k's $\Gamma(k - n + 1)$ is infinite. However, while it is true that the gamma function becomes infinite as its argument approaches $0, -1, -2, \dots$, it is not rigorous to say that it is infinite *at* those points; it is simply undefined there. Here, we ask you to verify that the $k = 0, 1, \dots, n-1$ terms are zero so that the correct lower limit in (17) is, indeed, $k = n$. For definiteness, let $\nu = 3$ and $r = -\nu = -3$. (You should then be able to generalize the result for the case of any positive integer ν, but we do not ask you to do that; $\nu = 3$ will suffice.) HINT: Rather than work from (17), go back to the formulas (4a,b,c).

11. It was stated, below (26), that J_ν and Y_ν are LI. Prove that claim. HINT: Use (25) for $\nu = n$ and (26) for $\nu \neq n$.

12. Each differential equation is given on $0 < x < \infty$. Use (50) to obtain a general solution. Gamma functions that appear need not be evaluated.

(a) $y'' + 4x^2 y = 0$ (b) $xy'' - 2y' - x^2 y = 0$
(c) $xy'' - 2y' + xy = 0$ (d) $4y'' + 9xy = 0$
(e) $y'' + \sqrt[3]{x}\, y = 0$ (f) $y'' - xy = 0$
(g) $xy'' + 3y' - xy = 0$ (h) $4xy'' + y = 0$
(i) $4xy'' + 2y' + xy = 0$ (j) $xy'' + 3y' - 4xy = 0$
(k) $xy'' + y' - 9x^2 y = 0$ (l) $y'' + xy = 0$

13. (a)–(l) Solve the corresponding problems in Exercise 12, this time using computer software.

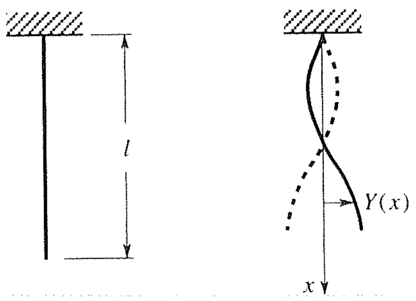
14. (a) Use (50) to find a general solution of

$$xy'' + 3y' + 9xy = 0. \quad (0 < x < \infty)$$

(b) Find a particular solution satisfying the boundary conditions $y(0) = 6,\ y'(0) = 0$.
(c) Show that there is *no* particular solution satisfying the initial conditions $y(0) = 6,\ y'(0) = 2$. Does that result contradict the result stated in Theorem 3.3.1? Explain.

15. Use (50) to solve $y'' + 4y = 0$, and show that your solution agrees with the known elementary solution. You may use any results given in these exercises.

16. (*Lateral vibration of hanging rope*) Consider a flexible rope or chain that hangs from the ceiling under the sole action of gravity (see the accompanying sketch). If we pull the rope to one side and let go, it will oscillate from side to side in a complicated pattern which amounts to a superposition of many different modes, each having a specific shape $Y(x)$ and temporal frequency ω. It can be shown (from Newton's second law of motion) that each shape $Y(x)$ is governed by the differential equation



$$[\rho g(l - x)Y']' + \rho\omega^2 Y = 0, \quad (0 < x < l) \tag{16.1}$$

where ρ is the mass per unit length and g is the acceleration of gravity.

(a) Derive the general solution

$$Y(x) = AJ_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right) + BY_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right) \tag{16.2}$$

of (16.1). HINT: It may help to first make the change of variables $l - x = \xi$. NOTE: Observe from (16.2) that the displacement Y will be unbounded at the free end $x = l$ because of the logarithmic singularity in Y_0 when its argument is zero (namely, when $x = l$). Mathematically, that singularity can be traced to the vanishing of the coefficient $\rho g(l - x)$ in (16.1) at $x = l$, which vanishing introduces a regular singular point of (16.1) at $x = l$ and results in the logarithmic singularity in the solution (16.2). Physically, observe that the coefficient $\rho g(l - x)$ in (16.1) represents the tension in the rope. The greater the tension the smaller the displacement (as anyone who has strung a clothesline knows). Hence the vanishing of the tension $\rho g(l - x)$ at the free end leads to the mathematical possibility of unbounded displacements there. In posing suitable boundary conditions, it is appropriate to preclude such unbounded displacements there by prescribing the boundary condition that $Y(l)$ be bounded; that is, a "boundedness condition." Imposing that condition implies that $B = 0$, so that the solution (16.2) reduces to $Y(x) = AJ_0\left(\frac{2\omega}{\sqrt{g}}\sqrt{l - x}\right)$.

(b) As a second boundary condition, set $Y(0) = 0$. That condition does not lead to the evaluation of A (which remains arbitrary); rather, it permits us to find the allowable temporal frequencies ω. If the first three zeros of $J_0(x)$ are $x = 2.405$, 5.520, and 8.654, evaluate the first three frequencies ω (in terms of g and l) and the corresponding mode shapes $Y(x)$ (to within the arbitrary scale factor A). Sketch those mode shapes by hand over $0 < x < l$.

(c) Use computer software to obtain the zeros quoted above (2.405, 5.520, 8.654), and to obtain computer plots of the three mode shapes. (Set $A = 1$, say.)

Chapter 4 Review

In this chapter we present methods for the solution of second-order homogeneous differential equations with nonconstant coefficients.

The most important general results are Theorems 4.2.4 and 4.3.1, which guarantee specific forms of series solutions about ordinary and regular singular points, respectively. About an ordinary point one can find two LI power series solutions and hence the general solution. About a regular singular point, say $x = 0$, one can find two LI solutions, by the method of Frobenius, in terms of power series and power series modified by the multiplicative factors $|x|^r$ and/or $\ln|x|$, where r is found by solving a quadratic equation known as the *indicial equation*. The combination of these forms is dictated by whether the roots r are repeated or distinct and, if distinct, whether they differ by an integer or not. Note that the $|x|^r$ and $\ln|x|$ factors introduce singularities in the solutions (unless r is a nonnegative integer).

Besides these general results, we meet these special functions:

Exponential integral (Section 4.3): $E_1(x) = \displaystyle\int_x^\infty \frac{e^{-t}}{t}\,dt, \quad (x > 0)$

Gamma function (Section 4.5): $\Gamma(x) = \displaystyle\int_0^\infty t^{x-1}e^{-t}\,dt, \quad (x > 0)$

and study these important differential equations and find solutions for them:

Legendre equation (Section 4.4): $(1 - x^2)y'' - 2xy' + \lambda y = 0$

Solutions that are bounded on $-1 \le x \le 1$ exist only if $\lambda = n(n+1)$ for $n = 0, 1, 2, \dots$, and these are the Legendre polynomials $P_n(x)$:

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1), \dots.$$

Bessel equation (Section 4.6): $x^2y'' + xy' + \left(x^2 - \nu^2\right)y = 0$

$$\text{General solution:} \qquad y(x) = \begin{cases} AJ_\nu(x) + BY_\nu(x) \\ CH_\nu^{(1)}(x) + DH_\nu^{(2)}(x) \end{cases}$$

where J_ν, Y_ν are Bessel functions of the first and second kind, respectively, of order ν, and $H_\nu^{(1)}, H_\nu^{(2)}$ are the Hankel functions of the first and second kind, respectively, of order ν.

Modified Bessel equation (Section 4.6): $x^2y'' + xy' + \left(-x^2 - \nu^2\right)y = 0$
For brevity, we consider only the case where $\nu = n$ is an integer.

$$\text{General solution:} \qquad y(x) = AI_n(x) + BK_n(x),$$

where I_n, K_n are modified Bessel functions of the first and second kinds, respectively, of order n.

NOTE: We suggest that to place the many Bessel function results in perspective it is helpful to see the Bessel equation $x^2y'' + xy' + (x^2 - \nu^2)y = 0$ and the modified Bessel equation $x^2y'' + xy' + (-x^2 - \nu^2)y = 0$ as analogous to the harmonic oscillator equation $y'' + y = 0$ and the "modified harmonic oscillator equation" $y'' - y = 0$. For instance, the oscillatory $J_\nu(x)$ and $Y_\nu(x)$ solutions of the Bessel equation are analogous to the oscillatory $\cos x$ and $\sin x$ solutions of $y'' + y = 0$, and the complex Hankel function solutions $H_\nu^{(1)}(x)$ and $H_\nu^{(2)}(x)$ are analogous to the complex e^{ix} and e^{-ix} solutions. Similarly, the nonoscillatory $I_\nu(x)$ and $K_\nu(x)$ solutions of the modified Bessel equation are analogous to the nonoscillatory e^x and e^{-x} solutions of the equation $y'' - y = 0$.

Equations reducible to Bessel equations (Section 4.6): The equation

$$\frac{d}{dx}\left(x^a \frac{dy}{dx}\right) + bx^c y = 0,$$

where a, b, c are real numbers, has solutions

$$y(x) = x^{\nu/\alpha} Z_{|\nu|}\left(\alpha\sqrt{|b|}\, x^{1/\alpha}\right),$$

where

$$\alpha = \frac{2}{c - a + 2}, \qquad \nu = \frac{1 - a}{c - a + 2}.$$

$Z_{|\nu|}$ denotes $J_{|\nu|}$ and $Y_{|\nu|}$ if $b > 0$, and $I_{|\nu|}$ and $K_{|\nu|}$ if $b < 0$.

Chapter 5

Laplace Transform

5.1 Introduction

The Laplace transform is an example of an **integral transform**, namely, a relation of the form

$$F(s) = \int_a^b K(t, s)\, f(t)\, dt, \tag{1}$$

which transforms a given function $f(t)$ into another function $F(s)$; $K(t, s)$ is called the **kernel** of the transform, and $F(s)$ is known as the **transform** of $f(t)$. Thus, whereas a function sends one number into another [for example, the function $f(x) = x^2$ sends the point $x = 3$ on an x axis into the point $f = 9$ on an f axis], (1) sends one function into another, namely, it sends $f(t)$ into $F(s)$. Probably the most well known integral transform is the **Laplace transform**, where $a = 0$, $b = \infty$, and $K(t, s) = e^{-st}$. In that case (1) takes the form

$$\boxed{F(s) = \int_0^\infty f(t)\, e^{-st}\, dt.} \tag{2}$$

The parameter s can be complex, but we limit it to real values in this chapter. Besides the notation $F(s)$ used in (2), the Laplace transform of $f(t)$ is also denoted as $L\{f(t)\}$ or as $\overline{f}(s)$, and in a given application we will use whichever of these three notations seems best.

The basic idea behind any transform is that the given problem can be solved more readily in the "transform domain." To illustrate, consider the use of the natural logarithm in numerical calculation. While the addition of two numbers is arithmetically simple, their multiplication can be quite laborious; for example, try working out 2.761359×8.247504 by hand. Thus, given two positive numbers u and v, suppose we wish to compute their product $y = uv$. Taking the logarithm of both sides gives $\ln y = \ln uv$. But $\ln uv = \ln u + \ln v$, so we have $\ln y = \ln u + \ln v$. Thus, whereas the original problem was one of multiplication, the problem in the "transform domain" is merely one of addition. The idea, then, is to look up $\ln u$ and $\ln v$

in a table and to add these two values. With the sum in hand, we again enter the table, this time using it in the reverse direction to find the antilog, y. (Of course, with pocket calculators and computers available logarithm tables are no longer needed, as they were fifty years ago, but the transform nature of the logarithm remains the same, whether we use tables or not.)

Similarly, the logarithm reduces exponentiation to multiplication since if $y = u^v$, then $\ln y = \ln(u^v) = v \ln u$, and it reduces division to subraction.

Analogously, given a linear ordinary differential equation with constant coefficients, we see that if we take a Laplace transform of all terms in the equation then we obtain a linear algebraic equation on the transform $X(s)$ of the unknown function $x(t)$. That equation can be solved for $X(s)$ by simple algebra and the solution $x(t)$ obtained from a Laplace transform table. The method is especially attractive for nonhomogeneous differential equations with forcing functions which are step functions or impulse functions; we study those cases in Section 5.5.

Observe that we have departed from our earlier usage of x as the independent variable. Here we use t and consider the interval $0 \leq t < \infty$ because in most (though not all) applications of the Laplace transform the independent variable is the time t, with $0 \leq t < \infty$.

A brief outline of this chapter follows:

5.2 *Calculation of the Transform.* In this section we study the existence of the transform, and its calculation.

5.3 *Properties of the Transform.* Three properties of the Laplace transform are discussed: linearity of the transform and its inverse, the transform of derivatives, and the convolution theorem. These are crucial in the application of the method to the solution of ordinary differential equations, homogeneous or not.

5.4 *Application to the Solution of Differential Equations.* Here, we demonstrate the principal application of the Laplace transform, namely, to the solution of linear ordinary differential equations.

5.5 *Discontinuous Forcing Functions; Heaviside Step Function.* Discontinuous forcing functions are common in engineering and science. In this section we introduce the Heaviside step function and demonstrate its use.

5.6 *Impulsive Forcing Function; Dirac Impulse Function.* Likewise common are impulsive forcing functions such as the force imparted to a mass by a hammer blow. In this section we introduce the Dirac delta function to model such impulsive actions.

5.7 *Additional Properties.* There are numerous useful properties of the transform beyond the three discussed in Section 5.3. A number of these are given here, as a sequence of theorems.

5.2 Calculation of the Transform

The first question to address is whether the transform $F(s)$ of a given function $f(t)$ exists – that is, whether the integral

$$F(s) = \int_0^\infty f(t)\, e^{-st}\, dt \tag{1}$$

converges. Before giving an existence theorem, we define two terms.

First, we say that $f(t)$ is of **exponential order** as $t \to \infty$ if there exist real constants K, c, and T such that

$$|f(t)| \leq Ke^{ct} \tag{2}$$

for all $t \geq T$. That is, the set of functions of exponential order is the set of functions that do not grow faster than exponentially, which includes most functions of engineering interest.

EXAMPLE 1. Is $f(t) = \sin t$ of exponential order? Yes: $|\sin t| \leq 1$ for all t, so (2) holds with $K = 1$, $c = 0$, and $T = 0$. Of course, these values are not uniquely chosen for (2) holds also with $K = 7$, $c = 12$, and $T = 100$, for instance. ∎

EXAMPLE 2. Is $f(t) = t^2$ of exponential order? l'Hôpital's rule gives

$$\lim_{t\to\infty} \frac{t^2}{e^{ct}} = \lim_{t\to\infty} \frac{2t}{c\, e^{ct}} = \lim_{t\to\infty} \frac{2}{c^2\, e^{ct}} = 0$$

if $c > 0$. Choose $c = 1$, say. Then, from the definition of limit, there must be a T such that $t^2/e^{ct} \leq 0.06$, say, for all $t \geq T$. Thus, $|f(t)| = t^2 \leq 0.06\, e^t$ for all $t \geq T$, hence $f(t)$ is of exponential order. ∎

On the other hand, the function $f(t) = e^{t^2}$ is not of exponential order because

$$\lim_{t\to\infty} \frac{e^{t^2}}{e^{ct}} = \lim_{t\to\infty} e^{t^2 - ct} = \infty, \tag{3}$$

no matter how large c is.

We say that $f(t)$ is **piecewise continuous** on $a \leq t \leq b$ if there exist a finite number of points $t_1, t_2, \ldots, t_N$ such that $f(t)$ is continuous on each open subinterval $a < t < t_1, t_1 < t < t_2, \ldots, t_N < t < b$, and has a finite limit as t approaches each endpoint from the interior of that subinterval. For instance, the function $f(t)$ shown in Fig. 1 is piecewise continuous on the interval $0 \leq t \leq 4$. The values of f *at* the endpoints $a, t_1, t_2, \ldots, b$ are not relevant to whether or not f is piecewise continuous; hence we have not even indicated those values in Fig. 1. For instance, the limit of f as t tends to 2 from the left exists and is 5, and the limit of f as t tends to 2 from the right exists and is 10, so the value of f at $t = 2$ does not matter. Thus, piecewise continuity allows for the presence of jump discontinuities.

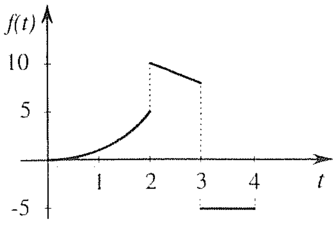


Figure 1. Piecewise continuity.

We can now provide a theorem that gives sufficient conditions on $f(t)$ for the existence of its Laplace transform $F(s)$.

THEOREM 5.2.1 *Existence of the Laplace Transform*
Let $f(t)$ satisfy these conditions: **(i)** $f(t)$ is piecewise continuous on $0 \leq t \leq A$, for every $A > 0$, and **(ii)** $f(t)$ is of exponential order as $t \to \infty$, so that there exist real constants K, c, and T such that $|f(t)| \leq Ke^{ct}$ for all $t \geq T$. Then the Laplace transform of $f(t)$, namely, $F(s)$ given by (1) exists for all $s > c$.

Proof: We need to show only that the singular integral in (1) is convergent. Breaking it up as

$$\int_0^\infty f(t)\, e^{-st}\, dt = \int_0^T f(t)\, e^{-st}\, dt + \int_T^\infty f(t)\, e^{-st}\, dt, \tag{4}$$

the first integral on the right exists since the integrand is piecewise continuous on the finite interval $0 < t < T$. In the second integral, $|f(t)e^{-st}| = |f(t)|e^{-st} \leq Ke^{-(s-c)t}$. Now, $\int_T^\infty Ke^{-(s-c)t}\, dt$ is convergent for $s > c$, so $\int_T^\infty f(t)\, e^{-st}\, dt$ is absolutely convergent – hence, by Theorem 4.5.3, convergent. ■

Being thus assured by Theorem 5.2.1 that the transform $F(s)$ exists for a large and useful class of functions, we proceed to illustrate the evaluation of $F(s)$ for several elementary functions, say $f(t) = 1,\ e^{at},\ \sin at$, where a is a real number, and $1/\sqrt{t}$.

EXAMPLE 3. If $f(t) = 1$, then the conditions of Theorem 5.2.1 are met for any $c \geq 0$, so according to Theorem 5.2.1, $F(s)$ should exist for all $s > 0$. Let us see.

$$F(s) = \int_0^\infty e^{-st}\, dt = \lim_{B\to\infty} \left. \frac{e^{-st}}{-s} \right|_0^B = \frac{1}{s}, \tag{5}$$

where the limit does indeed exist for all $s > 0$. Such restriction on s will cause no difficulty in applications. ■

EXAMPLE 4. If $f(t) = e^{at}$, the conditions of Theorem 5.2.1 are met for any $c \geq a$ so according to the theorem, $F(s)$ should exist for all $s > a$. In fact,

$$F(s) = \int_0^\infty e^{at}\, e^{-st}\, dt = \lim_{B\to\infty} \left. \frac{e^{-(s-a)t}}{-(s-a)} \right|_0^B = \frac{1}{s-a}, \tag{6}$$

where the limit does indeed exist for all $s > a$. ■

EXAMPLE 5. If $f(t) = \sin at$, then the conditions of Theorem 5.2.1 are met for any $c \geq 0$ so $F(s)$ should exist for all $s > 0$. In fact, integrating by parts twice gives

$$F(s) = \int_0^\infty \sin at\, e^{-st}\, dt$$

$$
\begin{aligned}
&= \lim_{B\to\infty}\left[\left(\sin at\frac{e^{-st}}{-s} + \frac{a}{s}\cos at\frac{e^{-st}}{-s}\right)\Bigg|_0^B - \frac{a^2}{s^2}\int_0^B \sin at\, e^{-st}\,dt\right] \\
&= (0-0) + \left(0 - \frac{a}{-s^2}\right) - \frac{a^2}{s^2}F(s), \qquad (7)
\end{aligned}
$$

where the limit exists if $s > 0$. The latter can be solved for $F(s)$, and gives

$$F(s) = \frac{a}{s^2+a^2} \qquad (s > 0) \qquad (8)$$

as the transform of $\sin at$.

COMMENT. An alternative approach, which requires a knowledge of the algebra of complex numbers (Section 21.2), is as follows:

$$
\begin{aligned}
\int_0^\infty \sin at\, e^{-st}\,dt &= \int_0^\infty \left(\operatorname{Im} e^{iat}\right) e^{-st}\,dt \\
&= \operatorname{Im}\int_0^\infty e^{-(s-ia)t}\,dt = \operatorname{Im}\left[\lim_{B\to\infty}\frac{e^{-(s-ia)t}}{-(s-ia)}\Bigg|_0^B\right] \\
&= \operatorname{Im}\frac{1}{s-ia} = \operatorname{Im}\frac{1}{s-ia}\frac{s+ia}{s+ia} = \frac{a}{s^2+a^2}, \qquad (9)
\end{aligned}
$$

as before, where the fourth equality follows because

$$\left|e^{-(s-ia)B}\right| = \left|e^{-sB}\right|\left|e^{iaB}\right| = e^{-sB} \to 0 \qquad (10)$$

as $B \to \infty$, if $s > 0$. In (10) we have used the fact that $|e^{iaB}| = |\cos aB + i\,\sin aB| = \sqrt{\cos^2 aB + \sin^2 aB} = 1$. ∎

EXAMPLE 6. If $f(t) = 1/\sqrt{t}$, then

$$F(s) = \int_0^\infty t^{-1/2}\,e^{-st}\,dt = \int_0^\infty \sqrt{\frac{s}{\tau}}\,e^{-\tau}\,\frac{d\tau}{s} = \frac{1}{\sqrt{s}}\int_0^\infty \tau^{-1/2}\,e^{-\tau}\,d\tau, \qquad (11)$$

where we have used the substitution $st = \tau$. Having studied the gamma function in Section 4.5, we see that the final integral is $\Gamma(1/2) = \sqrt{\pi}$ so

$$F(s) = \sqrt{\frac{\pi}{s}}, \qquad (12)$$

for $s > 0$ (Exercise 4). Observe that $f(t) = 1/\sqrt{t}$ does not satisfy the conditions of Theorem 5.2.1 because it is not piecewise continuous on $0 \le t < \infty$ since it does not have a limit as $t \to 0$. Nonetheless, the singularity at $t = 0$ is not strong enough to cause divergence of the integral in (11), and hence the transform exists. Thus, remember that all we need is convergence of the integral in (1); the conditions in Theorem 5.2.1 are sufficient, not necessary. ∎

From these examples we could begin to construct a Laplace transform table, with $f(t)$ in one column and its transform $F(s)$ in another. Such a table is supplied in Appendix C. More extensive ones are available,* and one can also obtain transforms and their inverses directly using computer software.

Tables can be used in either direction. For example, just as the transform of e^{at} is $1/(s-a)$, it is also true that the unique function whose transform is $1/(s-a)$ is e^{at}. Operationally, we say that

$$L\{e^{at}\} = \frac{1}{s-a} \quad \text{and} \quad L^{-1}\left\{\frac{1}{s-a}\right\} = e^{at}, \tag{13}$$

where L is the **Laplace transform operator** defined by

$$\boxed{L\{f(t)\} = \int_0^\infty f(t)\,e^{-st}\,dt,} \tag{14}$$

and L^{-1} is the **inverse Laplace transform operator**. It turns out that L^{-1} is, like L, an integral operator, namely

$$\boxed{L^{-1}\{F(s)\} = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)\,e^{st}\,ds,} \tag{15}$$

where γ is a sufficiently positive real number. The latter is an integration in a complex s plane, and to carry out such integrations one needs to study the complex integral calculus. If, for instance, we would put $1/(s-a)$ into the integrand in (15), for $F(s)$ and carry out the integration, we would obtain e^{at}. We will return to (15) near the end of this text, when we study the complex integral calculus, but we will not use it in our present discussion; instead, we will rely on tables (and computer software) to obtain inverses. In fact, there are entire books on the Laplace transform that do not even contain the inversion formula (15). Our purpose in presenting it here is to show that the inverse operator is, like L, an integral operator, and to close the operational "loop:" $L\{f(t)\} = F(s)$, and then $L^{-1}\{F(s)\} = f(t)$.

What can we say about the existence and uniqueness of the inverse transform? Although we do not need to go into them here, there are conditions that $F(s)$ must satisfy if the inversion integral, in (15), is to exist, to converge. Thus, if one writes a function $F(s)$ at random, it may not have an inverse; there may be no function $f(t)$ whose transform is that particular $F(s)$. For instance, there is no function $f(t)$ whose transform is s^2 because its inversion integral is divergent. But suppose that we can, indeed, find an inverse for a given $F(s)$. Is that inverse necessarily unique; might there be more than one function $f(t)$ having the same transform? Strictly speaking, the answer is always yes. For instance, not only does the function $f(t) = 1$ have the transform $F(s) = 1/s$ (as found in Example 3), but so does the function

$$g(t) = \begin{cases} 1, & 0 \le t < 3, \quad 3 < t < \infty \\ 500, & t = 3 \end{cases}$$

*See, for example, A. Erdèlyi (ed.), *Tables of Integral Transforms*, Vol. 1 (New York: McGraw-Hill, 1954).

have the transform $G(s) = 1$ because the integrands in $\int_0^\infty g(t)\, e^{-st}\, dt$ and $\int_0^\infty f(t)\, e^{-st}\, dt$ differ only at the single point $t = 3$. Since there is no area under a single point (of finite height), $G(s)$ and $F(s)$ are identical: $G(s) = F(s) = 1/s$.

Clearly, one can construct an *infinite* number of functions, each having $1/s$ as its transform, but in a practical sense these functions differ only superficially. In fact, it is known from **Lerch's theorem*** that the inverse transform is unique to within an additive *null function*, a function $N(t)$ such that $\int_0^T N(t)\, dt = 0$ for every $T > 0$, so we can be content that the inverse transform is essentially unique.

Closure. Theorem 5.2.1 guarantees the existence of the Laplace transform of a given function $f(t)$, subject to the (sufficient but not necessary) conditions that f be piecewise continuous on $0 \leq t < \infty$ and of exponential order as $t \to \infty$. We proceed to demonstrate the evaluation of the transforms of several simple functions, and discuss the building up of a transform table. Regarding the use of such a table, one needs to know whether the inverse transform found in the table is necessarily unique, and we use Lerch's theorem to show that for practical purposes we can, indeed, regard inverses as unique.

Computer software. On *Maple*, the **laplace** and **invlaplace** commands give transforms and inverses, respectively, provided that we enter readlib(laplace) first. To illustrate, the commands

```
readlib(laplace) :
laplace (1 + t^(-1/2) = x(t), t, s);
```

give the transform of $1 + t^{-1/2}$ as

$$\frac{1}{s} + \frac{\sqrt{\pi}}{\sqrt{s}} \tag{16}$$

and the command

```
invlaplace (a/(s^2 + a^2), s, t);          (17)
```

gives the inverse transform of $a/(s^2 + a^2)$ as $\sin at$.

EXERCISES 5.2

1. Show whether or not the given function is of exponential order. If it is, determine a suitable set of values for K, c, and T in (2).

(a) $5e^{4t}$ (b) $-10e^{-5t}$ (c) $\sinh 2t$
(d) $\cosh 3t$ (e) $\sinh t^2$ (f) $e^{4t} \sin t$
(g) $\cos t^3$ (h) t^{100} (i) $1/(t + 2)$
(j) $\cosh t^4$ (k) $6t + e^t \cos t$ (l) t^{1000}

2. If $f(t)$ is of exponential order, does it follow that df/dt is too? HINT: Consider $f(t) = \sin e^{t^2}$.

*See, for example, D. V. Widder, *The Laplace Transform* (Princeton, NJ: Princeton University Press, 1941).

3. If $f(t)$ and $g(t)$ are each of exponential order, does it follow that $f(g(t))$ is too? HINT: Consider the case where $f(t) = e^t$ and $g(t) = t^2$.

4. In Example 6 we state that the result (12) holds if $s > 0$. Show why that condition is needed.

5. Does $L\{t^{-3/2}\}$ exist? Explain.

6. Does $L\{t^{-2/3}\}$ exist? Explain.

7. Derive $L\{\cos at\}$ two ways: using integration by parts and using the fact that $\cos at = \operatorname{Re} e^{iat}$. (See Example 5.)

8. Derive $L\{e^{at} \sin bt\}$ two ways: using integration by parts and using $\sin bt = \operatorname{Im} e^{ibt}$.

9. Derive $L\{e^{at} \cos bt\}$ two ways: using integration by parts and using $\cos bt = \operatorname{Re} e^{ibt}$.

10. Derive by integration the Laplace transform for each of the following entries in Appendix C:

(a) entry 5 (b) entry 6 (c) entry 7 (d) entry 8

11. Derive entry 11 in Appendix C two ways:

(a) by writing the transform as $\frac{1}{2i}\int_0^\infty t\,e^{-(s-ia)t}\,dt - \frac{1}{2i}\int_0^\infty t\,e^{-(s+ia)t}\,dt$ and using integration by parts or, if you prefer, by writing the transform as the single integral $\operatorname{Im}\int_0^\infty t\,e^{-(s-ia)t}\,dt$ and using integration by parts;

(b) by differentiating both sides of the known transform

$$\int_0^\infty \sin at\, e^{-st}\,dt = \frac{a}{s^2+a^2}$$

(which we derived in Example 5) with respect to s, assuming the validity of the interchange

$$\frac{d}{ds}\int_0^\infty \sin at\, e^{-st}\,dt = \int_0^\infty \frac{d}{ds}\left(\sin at\, e^{-st}\right)dt$$

in the order of integration and differentiation.

12. Use the idea in Exercise 11(b) to derive

(a) entry (12) from entry (4) (b) entry (7) from entry (1)
(c) entry (13) from entry (5) (d) entry (14) from entry (6)
(e) entry (15) from entry (2)

13. Show that $L\{e^{at}\} = 1/(s-a)$ holds even if $a = \operatorname{Re} a + i \operatorname{Im} a$ is complex, provided that $s > \operatorname{Re} a$.

14. Use computer software to verify the given entry in Appendix C in both directions. That is, show that the transform of $f(t)$ is $F(s)$, and also show that the inverse of $F(s)$ is $f(t)$.

(a) 1–3 (b) 4–7 (c) 8–10 (d) 11–13
(e) 14–16 (f) 17–19 (g) 20–22

5.3 Properties of the Transform

When we studied the integral calculus we might have evaluated a few simple integrals directly from the definition of the Riemann integral, but for the most part we learned how to evaluate integrals using a number of properties. For instance, we used linearity (whereby $\int_a^b [\alpha u(x) + \beta v(x)]\,dx = \alpha\int_a^b u(x)\,dx + \beta\int_a^b v(x)\,dx$ for any constants α, β, and any functions u, v, if the two integrals on the right exist), integration by parts, and the fundamental theorem of the calculus (which enabled us to generate a long list of integrals from an already known long list of derivatives). Our plan for the Laplace transform is not much different; we work out a handful of transforms by direct integration, and then rely on a variety of properties of the transform and its inverse to extend that list substantially. There are many such properties, but in this section we present only the handful that will be essential when we apply the Laplace transform method to the solution of differential equations, in the next section. Additional properties are discussed in the final section of this chapter.

We begin with the linearity property of the transform and its inverse.

THEOREM 5.3.1 *Linearity of the Transform*
If $u(t)$ and $v(t)$ are any two functions such that the transforms $L\{u(t)\}$ and $L\{v(t)\}$ both exist, then

$$\boxed{L\{\alpha u(t) + \beta v(t)\} = \alpha L\{u(t)\} + \beta L\{v(t)\}} \tag{1}$$

for any constants α, β.

Proof: We have

$$\begin{aligned} L\{\alpha u(t) + \beta v(t)\} &= \int_0^\infty [\alpha u(t) + \beta v(t)]\; e^{-st}\, dt \\ &= \lim_{B\to\infty} \int_0^B [\alpha u(t) + \beta v(t)]\; e^{-st}\, dt \\ &= \lim_{B\to\infty} \left[\alpha \int_0^B u(t)\, e^{-st}\, dt + \beta \int_0^B v(t)\, e^{-st}\, dt\right] \\ &= \alpha \lim_{B\to\infty} \int_0^B u(t)\, e^{-st}\, dt + \beta \lim_{B\to\infty} \int_0^B v(t)\, e^{-st}\, dt \\ &= \alpha \int_0^\infty u(t)\, e^{-st}\, dt + \beta \int_0^\infty v(t)\, e^{-st}\, dt \\ &= \alpha L\{u(t)\} + \beta L\{v(t)\}, \end{aligned} \tag{2}$$

where the third equality follows from the linearity property of Riemann integration, and the fourth equality amounts to the result, from the calculus, that $\lim[\alpha f(B) + \beta g(B)] = \alpha \lim f(B) + \beta \lim g(B)$ as $B \to B_0$, if the latter two limits exist. ■

EXAMPLE 1. To evaluate the transform of $6 - 5e^{4t}$, for example, we need merely know the transforms of the simpler functions 1 and e^{4t} for $L\{6 - 5e^{4t}\} = 6L\{1\} - 5L\{e^{4t}\}$. Now, $L\{1\} = 1/s$ for $s > 0$, and $L\{e^{4t}\} = 1/(s-4)$ for $s > 4$ so

$$L\{6 - 5e^{4t}\} = 6\frac{1}{s} - 5\frac{1}{s-4} = \frac{s-24}{s(s-4)}$$

for $s > 4$. ■

THEOREM 5.3.2 *Linearity of the Inverse Transform*
For any $U(s)$ and $V(s)$ such that the inverse transforms $L^{-1}\{U(s)\} = u(t)$ and $L^{-1}\{V(s)\} = v(t)$ exist,

$$\boxed{L^{-1}\{\alpha U(s)+\beta V(s)\}=\alpha L^{-1}\{U(s)\}+\beta L^{-1}\{V(s)\}} \tag{3}$$

for any constants α, β.

Proof: Equation (3) follows either upon taking L^{-1} of both sides of (1) or from the linearity property of the integral in the inversion formula [equation (15) in Section 5.2]. ■

EXAMPLE 2. Asked to evaluate the inverse of $F(s) = 3/(s^2+3s-10)$, we turn to Appendix C but do not find this $F(s)$ in the column of transforms. However, we can simplify $F(s)$ by using partial fractions. Accordingly, we express

$$\begin{aligned}\frac{3}{s^2+3s-10} &= \frac{A}{s+5}+\frac{B}{s-2}\\ &= \frac{(A+B)s+(-2A+5B)}{s^2+3s-10}.\end{aligned} \tag{4}$$

To make the latter an identity we equate the coefficients of s^1 and s^0 in the numerators of the left- and right-hand sides: s^1 gives $0 = A+B$, and s^0 gives $3=-2A+5B$ so $A=3/7$ and $B=-3/7$. Then

$$\begin{aligned}L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= L^{-1}\left\{\frac{-3/7}{s+5}+\frac{3/7}{s-2}\right\}\\ &= -\frac{3}{7}L^{-1}\left\{\frac{1}{s+5}\right\}+\frac{3}{7}L^{-1}\left\{\frac{1}{s-2}\right\}\\ &= -\frac{3}{7}e^{-5t}+\frac{3}{7}e^{2t},\end{aligned} \tag{5}$$

where the second equality follows from Theorem 5.3.2, and the last equality follows from entry 2 in Appendix C.

COMMENT. Actually, we could have used entry 9 in Appendix C, which says that

$$L^{-1}=\left\{\frac{b}{(s-a)^2+b^2}\right\}=e^{at}\sin bt, \tag{6}$$

for if we equate $(s-a)^2+b^2 = s^2-2as+a^2+b^2$ to $s^2+3s-10$ we see that $a=-3/2$ and $b=\pm 7i/2$. Choose $b=+7i/2$, say ($b=-7i/2$ will give the same result). Then

$$\begin{aligned}L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= \frac{3}{7i/2}L^{-1}\left\{\frac{7i/2}{(s+3/2)^2+(7i/2)^2}\right\}\\ &= \frac{6}{7i}e^{-3t/2}\sin(7it/2) = \frac{6}{7i}e^{-3t/2}\,\frac{e^{i(7it/2)}-e^{-i(7it/2)}}{2i}\\ &= \frac{3}{7}\left(-e^{-5t}+e^{2t}\right),\end{aligned} \tag{7}$$

where the first equality is true by linearity and the second follows from (6). This result is the same as the one found above by partial fractions. This example illustrates the fact that we can often invert a given transform in more than one way. ▮

If we are going to apply the Laplace transform method to differential equations, we need to know how to take transforms of derivatives.

THEOREM 5.3.3 *Transform of the Derivative*
Let $f(t)$ be continuous and $f'(t)$ be piecewise continuous on $0 \leq t \leq t_0$ for every finite t_0, and let $f(t)$ be of exponential order as $t \to \infty$ so that there are constants K, c, T such that $|f(t)| \leq Ke^{ct}$ for all $t > T$. Then $L\{f'(t)\}$ exists for all $s > c$, and

$$\boxed{L\{f'(t)\} = s\,L\{f(t)\} - f(0).} \tag{8}$$

Proof: Since $L\{f'(t)\} = \lim_{B\to\infty} \int_0^B f'(t)\,e^{-st}\,dt$, consider the integral

$$\mathrm{I} = \int_0^B f'(t)\,e^{-st}\,dt = \int_0^{t_1} f'(t)\,e^{-st}\,dt + \cdots + \int_{t_n}^B f'(t)\,e^{-st}\,dt, \tag{9}$$

where $t_1, \ldots, t_n$ are the points, in $0 < t < B$, at which f' is discontinuous. Integrating by parts gives

$$\begin{aligned}\mathrm{I} = {} & f(t)\,e^{-st}\Big|_0^{t_1} + \cdots + f(t)\,e^{-st}\Big|_{t_n}^B \\ & + s\int_0^{t_1} f(t)\,e^{-st}\,dt + \cdots + s\int_{t_n}^B f(t)\,e^{-st}\,dt.\end{aligned} \tag{10}$$

By virtue of the continuity of f, the boundary terms at $t_1, \ldots, t_n$ cancel in pairs so that, after recombining the integrals in (10), we have

$$\mathrm{I} = f(B)\,e^{-sB} - f(0) + s\int_0^B f(t)\,e^{-st}\,dt. \tag{11}$$

Since f is of exponential order as $t \to \infty$ it follows that $f(B)\,e^{-sB} \to 0$ as $B \to \infty$. Thus,

$$\begin{aligned}L\{f'(t)\} &= \lim_{B\to\infty}\left[f(B)\,e^{-sB} - f(0) + s\int_0^B f(t)\,e^{-st}\,dt\right] \\ &= 0 - f(0) + s\,L\{f(t)\},\end{aligned} \tag{12}$$

as was to be proved. ▪

The foregoing result can be used to obtain the transforms of higher derivatives as well. For example, if $f'(t)$ satisfies the conditions imposed on f in Theorem 5.2.3, then replacement of f by f' in (8) gives

$$L\{f''\} = s\,L\{f'\} - f'(0) = s\,[s\,L\{f\} - f(0)] - f'(0).$$

If, besides f', f also satisfies the conditions of Theorem 5.3.3, so that the $L\{f\}$ term on the right side exists, then

$$L\{f''\} = s^2\,L\{f\} - s\,f(0) - f'(0). \tag{13}$$

Similarly,

$$L\{f'''\} = s^3\,L\{f\} - s^2\,f(0) - s\,f'(0) - f''(0), \tag{14}$$

if f'', f', and f satisfy the conditions of Theorem 5.3.3, and so on for the transforms of higher-order derivatives.

The last of the major properties of the Laplace tansform that we discuss in this section is the Laplace convolution theorem.

THEOREM 5.3.4 *Laplace Convolution Theorem*
If $L\{f(t)\} = F(s)$ and $L\{g(t)\} = G(s)$ both exist for $s > c$, then

$$\boxed{L^{-1}\{F(s)G(s)\} = \int_0^t f(\tau)\,g(t-\tau)\,d\tau} \tag{15}$$

or, equivalently,

$$\boxed{L\left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} = F(s)G(s)} \tag{16}$$

for $s > c$.

Proof: Since (15) and (16) are equivalent statements, it suffices to prove just one, say (16). By definition,

$$L\left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} = \int_0^\infty \left\{\int_0^t f(\tau)\,g(t-\tau)\,d\tau\right\} e^{-st}\,dt. \tag{17}$$

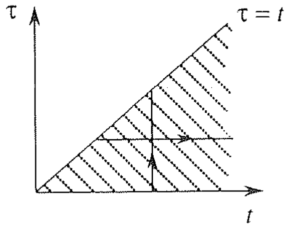


Figure 1. Region of integration.

Regarding the latter as an iterated integral over a 45° wedge in a τ, t plane as shown in Fig. 1, let us invert the order of integration in (17). Recalling from the calculus the equivalent notations

$$\int_c^d \left\{\int_a^b f(x,y)\,dx\right\} dy = \int_c^d \int_a^b f(x,y)\,dx\,dy = \int_c^d dy \int_a^b f(x,y)\,dx \tag{18}$$

for iterated integrals (where a, b, c, d are constants, say), inverting the order of integration in (17) gives

$$\begin{aligned} L\left\{\int_0^t f(\tau)\, g(t-\tau)\, d\tau\right\} &= \int_0^\infty \int_\tau^\infty f(\tau)\, g(t-\tau)\, e^{-st} dt\, d\tau \\ &= \int_0^\infty f(\tau)\, d\tau \int_\tau^\infty g(t-\tau)\, e^{-st}\, dt \\ &= \int_0^\infty f(\tau)\, d\tau \int_0^\infty g(\mu)\, e^{-s(\mu+\tau)}\, d\mu \\ &= \int_0^\infty f(\tau)\, e^{-s\tau}\, d\tau \int_0^\infty g(\mu)\, e^{-s\mu}\, d\mu. \end{aligned} \tag{19}$$

The last product is simply $F(s)$ times $G(s)$, so the theorem is proved. ■

The integral on the right side of (15) is called the **Laplace convolution** of f and g and is denoted as $f * g$. It too is a function of t:

$$\boxed{(f * g)(t) \equiv \int_0^t f(\tau)\, g(t-\tau)\, d\tau.} \tag{20}$$

CAUTION: Be sure to see that the inverse of the product, $L^{-1}\{F(s)G(s)\}$, is not simply the algebraic product of the inverses, $f(t)g(t)$; rather, it is (according to Theorem 5.3.4) their convolution, $(f * g)(t)$.

EXAMPLE 3. In Example 2 we inverted $F(s) = 3/(s^2 + 3s - 10)$ in two different ways. Let us now obtain the inverse by still another method, the convolution theorem:

$$\begin{aligned} L^{-1}\left\{\frac{3}{s^2+3s-10}\right\} &= 3L^{-1}\left\{\frac{1}{s-2}\,\frac{1}{s+5}\right\} = 3L^{-1}\left\{\frac{1}{s-2}\right\} * L^{-1}\left\{\frac{1}{s+5}\right\} \\ &= 3e^{2t} * e^{-5t} = 3\int_0^t e^{2\tau} e^{-5(t-\tau)}\, d\tau \\ &= \frac{3}{7}\left(e^{2t} - e^{-5t}\right), \end{aligned} \tag{21}$$

which is the same result as obtained in Example 2. ■

Observe that in equation (15) it surely doesn't matter if we write $F(s)G(s)$ or $G(s)F(s)$ because ordinary multiplication is commutative. Yet it is not clear that the results are the same, $\int_0^t f(\tau)\, g(t-\tau)\, d\tau$ in one case and $\int_0^t g(\tau)\, f(t-\tau)\, d\tau$ in the other. Nonetheless, these results are indeed the same, proof of which claim is left as an exercise. In fact, although the convolution is not an ordinary product it does share several of the properties of ordinary multiplication:

$$f * g = g * f, \qquad \text{(commutative)} \tag{22a}$$

$$f * (g * h) = (f * g) * h, \qquad \text{(associative)} \qquad (22b)$$
$$f * (g + h) = f * g + f * h, \qquad \text{(distributive)} \qquad (22c)$$
$$f * 0 = 0. \qquad (22d)$$

Closure. The properties studied in this section – linearity, the transform of a derivative, and the convolution theorem, should be thoroughly understood. All are used in the next section, where we use the Laplace transform method to solve differential equations. The convolution property, in particular, should be studied carefully.

The convolution theorem is useful in both directions. If we have a transform $H(s)$ that is difficult to invert, it may be possible to factor H as $F(s)G(s)$, where F and G are more easily inverted. If so, then $h(t)$ is given, according to (15), as the convolution of $f(t)$ and $g(t)$. Furthermore, we may need to find the transform of an integral that is in convolution form. If so, then the transform is given easily by (16).

Finally, we mention that convolution integrals arise in *hereditary* systems, systems whose behavior at time t depends not only on the state of the system at that instant but also on its past history. Examples occur in the study of viscoelasticity and population dynamics.

EXERCISES 5.3

1. Find the inverse of the given transform two different ways: using partial fractions and using the convolution theorem. Cite any entries used from Appendix C.

(a) $3/[s(s+8)]$ (b) $1/(3s^2+5s-2)$
(c) $1/(s^2-a^2)$ (d) $5/[(s+1)(3s+2)]$
(e) $1/(s^2+s)$ (f) $2/(2s^2-s-1)$

2.(a)–(f) Find the inverse of the corresponding transform in Exercise 1 using computer software.

3. Use entry 9 in Appendix C to evaluate the inverse of each. If necessary, use entry 10 as well. NOTE: See the Comment in Example 2.

(a) $1/(s^2+8s)$ (b) $1/(s^2-3s+3)$
(c) $1/(s^2-s)$ (d) $1/(s^2-s-2)$
(e) $s/(s^2-2s+2)$ (f) $(s+1)/(s^2+4s+6)$
(g) $(s+1)/(s^2-s)$ (h) $(2s-1)/(s^2-6s+5)$

4. Use (8) together with mathematical induction to verify the general formula $L\{f^{(n)}\} = s^n L\{f\} - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0)$, which is valid if $f^{(n-1)}, f^{(n-2)}, \ldots, f'$, and f satisfy the conditions of Theorem 5.3.3.

5. Prove

(a) equation (22a) (b) equation (22b)
(c) equation (22c) (d) equation (22d)

6. Prove that $L\{f * g * h\} = F(s)G(s)H(s)$ or, equivalently, that $L^{-1}\{F(s)G(s)H(s)\} = f * g * h$. NOTE: Does $f * g * h$ mean $(f * g) * h$ or $f * (g * h)$? According to the associative property (22b) it doesn't matter; they are equal.

7. To illustrate the result stated in Exercise 6, find the inverse of $1/s^3$ as $L^{-1}\left\{\frac{1}{s^3}\right\} = L^{-1}\left\{\frac{1}{s}\frac{1}{s}\frac{1}{s}\right\} = 1 * 1 * 1$, and show that the result agrees with that given directly in Appendix C.

8. Factoring $\frac{s}{(s^2+a^2)^2} = \frac{s}{s^2+a^2}\,\frac{1}{s^2+a^2}$, it follows from the convolution theorem and entries 3 and 4 of Appendix C that

$$L^{-1}\left\{\frac{s}{(s^2+a^2)^2}\right\} = \cos at * \frac{\sin at}{a}.$$

Evaluate this convolution and show that the result agrees with that given directly by entry 11.

9. Verify (8) and (13) directly, for each given $f(t)$, by working out the left- and right-hand sides and showing that they are equal. You may use the table in Appendix C to evaluate $L\{f''(t)\}$, $L\{f'(t)\}$, and $L\{f(t)\}$.

(a) e^{3t} (b) $e^{-4t}+2$ (c) t^2+5t-1
(d) $\sinh 4t$ (e) $\cosh 3t + 5t^6$ (f) $4t^{3/2}-\cos 2t$

10. Evaluate the transform of each:

(a) $\int_0^t e^{t-\tau}\sin 2\tau\, d\tau$ (b) $\int_0^t \cos 3(t-\tau)\, d\tau$
(c) $\int_0^t (t-\tau)^8 e^{-3\tau}\, d\tau$ (d) $\int_0^t \cosh 3(t-\tau)\, d\tau$
(e) $\int_0^t \dfrac{\sin\tau}{\sqrt{t-\tau}}\, d\tau$ (f) $\int_0^t \tau^{5.2}\sinh 4(t-\tau)\, d\tau$

11. We emphasized that $L^{-1}\{F(s)G(s)\}$ equals the convolution of f and g; in general, it does not merely equal the product $f(t)g(t)$. Show that $L^{-1}\{F(s)G(s)\} \neq f(t)g(t)$ for each given pair of functions.

(a) $f(t)=t,\ \ g(t)=e^t$ (b) $f(t)=\sin t,\ \ g(t)=4$
(c) $f(t)=t,\ \ g(t)=t^2$ (d) $f(t)=\cos t,\ \ g(t)=t+6$

5.4 Application to the Solution of Differential Equations

Our object, in this section, is to explain the use of the Laplace transform in solving linear constant-coefficient differential equations on the interval $0<t<\infty$, with initial conditions at $t=0$.

EXAMPLE 1. We've already studied the important case of the harmonic oscillator – both free and forced, both damped and undamped. Consider the undamped mechanical oscillator shown in Fig. 1, with a forcing function that is a constant: $f(t) = F_0$. Recall that the displacement $x(t)$ then satisfies the equation

$$mx'' + kx = f(t) = F_0. \tag{1}$$

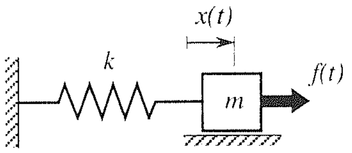


Figure 1. Mechanical oscillator.

Further, we assume the initial conditions $x(0)$ and $x'(0)$ are known.

To apply the Laplace transform, we transform equation (1). That is, we multiply each term in (1) by the Laplace kernel e^{-st} and integrate on t from 0 to ∞. Operationally, we use L to denote that step:

$$L\{mx''+kx\} = L\{F_0\}. \tag{2}$$

By the linearity of L (Theorem 5.3.1), we can rewrite (2) as

$$mL\{x''(t)\} + kL\{x(t)\} = F_0\, L\{1\}. \tag{3}$$

Recalling from Theorem 5.3.2 that $L\{x''(t)\} = s^2X(s) - sx(0) - x'(0)$, noting that $L\{x(t)\} \equiv X(s)$, and obtaining $L\{1\} = 1/s$ from Appendix C, (3) becomes

$$m\left[s^2X(s) - sx(0) - x'(0)\right] + kX(s) = F_0\,\frac{1}{s}. \tag{4}$$

The point to appreciate is that whereas in the t domain we had the linear *differential* equation (1) on $x(t)$, in the transform domain, or s domain, we now have the linear *algebraic* equation (4) on $X(s)$. The solution now amounts to solving (4) by simple algebra for $X(s)$. Doing so gives

$$X(s) = \frac{sx(0)+x'(0)}{s^2+\omega^2} + \frac{F_0}{ms\,(s^2+\omega^2)}, \tag{5}$$

where $\omega = \sqrt{k/m}$ is the natural frequency.

With the solving for $X(s)$ completed, we now invert (5) to obtain $x(t)$:

$$\begin{aligned} x(t) &= L^{-1}\left\{\frac{sx(0)+x'(0)}{s^2+\omega^2} + \frac{F_0}{ms\,(s^2+\omega^2)}\right\} \\ &= x(0)L^{-1}\left\{\frac{s}{s^2+\omega^2}\right\} + x'(0)L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} + \frac{F_0}{m}L^{-1}\left\{\frac{1}{s\,(s^2+\omega^2)}\right\}, \quad (6)\end{aligned}$$

where the second equality follows from the linearity of the L^{-1} operator (Theorem 5.3.2). Appendix C gives

$$L^{-1}\left\{\frac{s}{s^2+\omega^2}\right\} = \cos\omega t \qquad \text{and} \qquad L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} = \frac{\sin\omega t}{\omega}, \tag{7}$$

but the third inverse in (6) is not found in the table. We could evaluate it with the help of partial fractions, but it is easier to use the convolution theorem:

$$\begin{aligned} L^{-1}\left\{\frac{1}{s\,(s^2+\omega^2)}\right\} &= L^{-1}\left\{\frac{1}{s}\,\frac{1}{s^2+\omega^2}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} \\ &= 1 * \frac{\sin\omega t}{\omega} = \int_0^t (1)\left(\frac{\sin\omega(t-\tau)}{\omega}\right)d\tau = \frac{1-\cos\omega t}{\omega^2}, \quad (8)\end{aligned}$$

so (6), (7), and (8) give the desired particular solution as

$$x(t) = x(0)\,\cos\omega t + \frac{x'(0)}{\omega}\,\sin\omega t + \frac{F_0}{k}\,(1-\cos\omega t)\,. \tag{9}$$

For instance, if $x(0) = x'(0) = 0$, then $x(t) = (F_0/k)(1-\cos\omega t)$ as depicted in Fig. 2. Does it seem correct that the constant force F_0 should cause an oscillation? Yes, for imagine rotating the apparatus 90° so that the mass hangs down. Then we can think of F_0 as the downward gravitational force on m. In static equilibrium, the mass will hang down an amount $x = F_0/k$. If we release it from $x = 0$, it will fall and then oscillate about the equilibrium position $x = F_0/k$, as shown in Fig. 2.

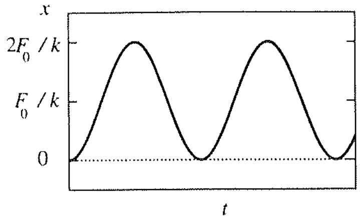


Figure 2. Release from rest.

COMMENT 1. Recall that $f * g = g * f$, so we can write the convolution integral either as $\int_0^t f(\tau)\,g(t-\tau)\,d\tau$ or as $\int_0^t g(\tau)\,f(t-\tau)\,d\tau$; that is, we can let the argument of f be τ and the argument of g be $t-\tau$, or vice versa, whichever we choose. In (8) we chose the τ argument for 1 and the $t-\tau$ argument for $(\sin\omega t)/\omega$. (Of course, if we change all the t's in 1 to τ's we still have 1 because there are no t's in 1.) Alternatively, we could have expressed the inverse in (8) as

$$\frac{\sin\omega t}{\omega} * 1 = \int_0^t \left(\frac{\sin\omega\tau}{\omega}\right)(1)\,d\tau = \frac{1-\cos\omega\tau}{\omega^2},$$

as obtained in (8).

COMMENT 2. Observe that (9) is the particular solution satisfying the initial conditions $x = x(0)$ and $x' = x'(0)$ at $t = 0$. If those quantities are not prescribed, we can replace them by arbitrary constants, and then (9) amounts to the general solution of $mx'' + kx =$

F_0. Thus, the method gives either a particular solution or a general solution, whichever is desired.

COMMENT 3. If, instead of the specific forcing function $f(t) = F_0$ we allow $f(t)$ to be an unspecified function, then we have, in place of (5),

$$X(s) = \frac{sx(0) + x'(0)}{s^2 + \omega^2} + \frac{F(s)}{m(s^2 + \omega^2)}, \tag{10}$$

and, in place of (9),

$$x(t) = x(0)\,\cos\omega t + \frac{x'(0)}{\omega}\,\sin\omega t + \frac{1}{m}L^{-1}\left\{\frac{F(s)}{s^2 + \omega^2}\right\}. \tag{11}$$

Using the convolution theorem to write

$$L^{-1}\left\{\frac{F(s)}{s^2 + \omega^2}\right\} = L^{-1}\left\{\frac{1}{s^2 + \omega^2}\right\} * L^{-1}\left\{F(s)\right\} = \frac{\sin\omega t}{\omega} * f(t), \tag{12}$$

gives

$$x(t) = x(0)\,\cos\omega t + \frac{x'(0)}{\omega}\,\sin\omega t + \frac{1}{m\omega}\int_0^t \sin\omega\tau\, f(t - \tau)\,d\tau \tag{13}$$

as the solution. ∎

With Example 1 completed, there are several observations that can be made about the method. First, consider the general second-order equation

$$x'' + ax' + bx = f(t), \tag{14}$$

where a, b are constants, although the following discussion applies to higher-order equations as well. If we solve (14) by the methods of Chapter 3, then we need both homogeneous and particular solutions. To find the homogeneous solution we need to factor the characteristic polynomial $\lambda^2 + a\lambda + b$ or, equivalently, to find the roots of the characteristic equation $\lambda^2 + a\lambda + b = 0$. Solving (14) by the Laplace transform instead, we obtain, and need to invert,

$$X(s) = \frac{(s + a)x(0) + x'(0)}{s^2 + as + b} + \frac{F(s)}{s^2 + as + b}. \tag{15}$$

Whether we invert these terms by partial fractions or by some other method, their inversion depends, essentially, on our being able to factor the $s^2 + as + b$ denominator. That polynomial is none other than the characteristic polynomial corresponding to (14). Thus, whether we solve (14) by seeking exponential solutions for the homogeneous equation and then seeking a particular solution, or by using the Laplace transform, we need to face up to the same task, finding the roots of the characteristic equation.

Second, observe that if we invert the $F(s)/(s^2 + as + b)$ term in (15) by the convolution theorem, then we convolve the inverse of $F(s)$, namely, $f(t)$, with the inverse of $1/(s^2 + as + b)$. Therefore, if we use the convolution theorem, then there is no need to evaluate the transform $F(s)$ of $f(t)$ when transforming the given differential equation.

Third, observe how the initial conditions become "built in," when we take the transform of the differential equation. Thus, there is no need to apply them at the end.

Fourth, recall that Laplace transforms come with restrictions on s. For instance, $L\{1\} = 1/s$ for $s > 0$. However, such restrictions in no way impede the solution steps in using the Laplace transform method, and once we invert $X(s)$ to obtain $x(t)$ they are no longer relevant.

Fifth, we need to realize that when we apply the Laplace transform method to a differential equation, we take the transform of the unknown and one or more of its derivatives, but since we don't yet know the solution we don't yet know whether or not these functions are transformable. The procedure, then, is to assume that they are transformable in order to proceed, and to verify that they are once the solution is in hand.

Finally, and most important, understand that the power of the Laplace transform, in solving linear constant-coefficient differential equations, is in its ability to convert such an equation to a linear algebraic equation on $X(s)$, which ability flows from the fact that the transform of $f'(t)$ is merely a multiple of $F(s)$ plus a constant (and therefore similarly for $f'', f''', \ldots$). Indeed, the transform $L\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt$ was *designed* so as to have this property. That is, the "kernel" e^{-st} was designed so as to imbue the transform with that property.

EXAMPLE 2. Solve the initial-value problem

$$y^{(iv)} - y = 0; \qquad y(0) = 1,\ y'(0) = y''(0) = y'''(0) = 0 \tag{16}$$

for $y(x)$. That the independent and dependent variables are x, y, rather than x, t, is immaterial to the application of the Laplace transform; the transform of $y(x)$ is now $Y(s) = \int_0^\infty y(x)\, e^{-sx}\, dx$. Taking the transform of (16) gives

$$\left[s^4 Y(s) - s^3 y(0) - s^2 y'(0) - s y''(0) - y'''(0)\right] - Y(s) = 0. \tag{17}$$

Putting the initial conditions into (17), and solving for $Y(s)$, gives

$$Y(s) = \frac{s^3}{s^4 - 1}. \tag{18}$$

To invert the latter, we can use partial fractions:

$$\begin{aligned} \frac{s^3}{s^4-1} &= \frac{A}{s+1} + \frac{B}{s-1} + \frac{C}{s+i} + \frac{D}{s-i} \\ &= \frac{(s-1)(s^2+1)A + (s+1)(s^2+1)B + (s-i)(s^2-1)C + (s+i)(s^2-1)D}{s^4-1}. \end{aligned} \tag{19}$$

Equating coefficients of like powers of s in the numerators gives the linear equations

$$\begin{aligned} s^3 &: \quad 1 = A + B + C + D, \\ s^2 &: \quad 0 = -A + B - iC + iD, \\ s &: \quad 0 = A + B - C - D, \\ 1 &: \quad 0 = -A + B + iC - iD, \end{aligned}$$

solution of which (for instance by Gauss elimination) gives $A = B = C = D = 1/4$. Thus,

$$\begin{aligned} y(x) &= \frac{1}{4}e^{-x} + \frac{1}{4}e^{x} + \frac{1}{4}e^{-ix} + \frac{1}{4}e^{ix} \\ &= \frac{1}{2}(\cosh x + \cos x) \end{aligned} \tag{20}$$

is the desired particular solution. ∎

EXAMPLE 3. Solve the first-order initial-value problem

$$x' + px = q(t); \qquad x(0) = x_0 \tag{21}$$

for $x(t)$, where p is a constant and $q(t)$ is any prescribed forcing function. Application of the Laplace transform gives

$$X(s) = \frac{x_0}{s+p} + \frac{Q(s)}{s+p} \tag{22}$$

and hence the particular solution

$$\begin{aligned} x(t) &= x_0 e^{-pt} + \int_0^t e^{-p(t-\tau)}\, q(\tau)\, d\tau \\ &= e^{-pt}\left[x_0 + \int_0^t q(\tau)\, e^{p\tau}\, d\tau\right]. \end{aligned} \tag{23}$$

COMMENT. Alternatively, let us begin by integrating the differential equation on t, from 0 to t:

$$x(t)\Big|_0^t + p\int_0^t x(\tau)\, d\tau = \int_0^t q(\tau)\, d\tau \tag{24}$$

or, since $x(0) = x_0$,

$$x(t) + p\int_0^t x(\tau)\, d\tau = x_0 + \int_0^t q(\tau)\, d\tau. \tag{25}$$

Of course, (25) is not the solution of the differential equation because the unknown $x(t)$ is under the integral sign. Thus, (25) is an example of an **integral equation**. Although we will not study integral equations systematically in this text, it will be useful to at least introduce them. Observe, first, that (25) is equivalent to both the differential equation and the initial condition, for they led to (25); conversely, the derivative of (25) gives back $x' + px = q(t)$

[if $q(t)$ is continuous], and putting $t = 0$ in (25) gives back $x(0) = x_0$. That is, unlike the differential equation, the integral equation version has the initial condition "built in."

Further, we can solve (25) by the Laplace transform conveniently because each integral is of convolution type: the first is $1 * x(t)$, and the second is $1 * q(t)$. Thus, taking a Laplace transform of (25), and noting that $L\{1 * x(t)\} = L\{1\}L\{x(t)\} = (1/s)X(s)$ and $L\{1 * q(t)\} = (1/s)Q(s)$, gives

$$X(s) + p\frac{1}{s}X(s) = \frac{x_0}{s} + \frac{1}{s}Q(s), \tag{26}$$

which, once again, gives (22) and hence the solution (23). ▮

Closure. In this section we describe the application of the Laplace transform to the solution of linear differential equations with constant coefficients, homogeneous or nonhomogeneous. In a sense, the method is of comparable difficulty to the solution methods studied in Chapter 3 in that one still needs to be able to factor the characteristic polynomial, which can be difficult if the equation is of high order. However, the Laplace transform method has a number of advantages. First, the method reduces a linear differential equation to a linear algebraic equation. Second, the hardest part, namely the inversion of the transform of the unknown, can often be accomplished with the help of tables or computer software, as well as with several additional theorems that are given in the final section of this chapter. Third, any initial conditions that are given become built in, in the process of taking the transform of the differential equation, so they do not need to be applied separately, at the end, as they were in Chapter 3.

We also saw, in the final example, that the Laplace transform is convenient to use in solving integral equations (equations in which the unknown function appears under an integral sign), provided that the integrals therein are Laplace convolution integrals; additional discussion of this idea is left for the exercises. In fact, it might be noted that the Laplace transform itself, $F(s) = \int_0^\infty f(t)\, e^{-st}\, dt$, is really an integral equation for $f(t)$ if $F(s)$ is known. Although that integral equation was studied by Laplace, it was *Simeon-Denis Poisson* (1781 – 1840) who discovered the solution $f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)\, e^{st}\, ds$, namely, the Laplace inversion formula. Poisson was one of the great nineteenth century analysts and a professor at the Ecole Polytechnique.

Also left for the exercises is discussion of the application of the method to a limited class of nonconstant differential equations.

EXERCISES 5.4

1. Use the Laplace transform to find the general solution, or the particular solution if initial conditions are given.

(a) $x' + 2x = 4t^2$
(b) $3x' + x = 6e^{2t}; \quad x(0) = 0$
(c) $x' - 6x = e^{-t}; \quad x(0) = 4$
(d) $x'' = 6t; \quad x(0) = 2,\ x'(0) = -1$
(e) $x'' + 5x' = 10$
(f) $x'' - x' = 1 + t + t^2$

(g) $x'' - 3x' + 2x = 0; \quad x(0) = 3,\ x'(0) = 1$
(h) $x'' - 4x' - 5x = 2 + e^{-t}; \quad x(0) = x'(0) = 0$
(i) $x'' - x' - 12x = t; \quad x(0) = -1,\ x'(0) = 0$
(j) $x'' + 6x' + 9x = 1; \quad x(0) = 0,\ x'(0) = -2$
(k) $x'' - 2x' + 2x = -2t; \quad x(0) = 0,\ x'(0) = -5$
(l) $x'' - 2x' + 3x = 5; \quad x(0) = 1,\ x'(0) = -1$
(m) $x''' - x'' + 2x' = t^2; \quad x(0) = 1,\ x'(0) = x''(0) = 0$
(n) $x''' + x'' - 2x' = 1 + e^t; \quad x(0) = x'(0) = x''(0) = 0$
(o) $x''' + 5x'' = t^4; \quad x(0) = x'(0) = 0,\ x''(0) = 1$
(p) $x''' + 3x'' + 3x' + x = e^{2t}$
(q) $x''' - x'' - x' + x = 0; \quad x(0) = 2,\ x'(0) = x''(0) = 0$
(r) $x^{(iv)} = 2\sin t$
(s) $x^{(iv)} + 3x''' = 0; \quad x(0) = x'(0) = 0,$
$x''(0) = x'''(0) = 3$
(t) $x^{(iv)} + 8x'' + 16x = 4$
(u) $x^{(iv)} - x = 1; \quad x(0) = x'(0) = x''(0) = 0,$
$x'''(0) = 4$
(v) $x^{(iv)} - x = 4; \quad x(0) = x'(0) = x''(0) = 1,$
$x'''(0) = 0$
(w) $x^{(iv)} - 16x = -32; \quad x(0) = 0,\ x'(0) = 2,$
$x''(0) = x'''(0) = 0$

2. (a) Show that for a constant-coefficient linear homogeneous differential equation of order n, the Laplace transform $X(s)$ of the solution $x(t)$ is necessarily of the form

$$X(s) = P(s)/Q(s), \tag{2.1}$$

where $Q(s)$ and $P(s)$ are polynomials in s, with Q of degree n and P of degree less than n.

(b) Show that if $Q(s) = 0$ has n distinct roots $r_1, \dots r_n$, then

$$x(t) = \sum_{k=1}^{n} \frac{P(r_k)}{Q'(r_k)} e^{r_k t}. \tag{2.2}$$

3. Our purpose, in this exercise, is to follow up on Example 3 in showing a connection between differential equations and integral equations, and in considering the solution of certain integral equations by the Laplace transform method.

(a) Convert the initial–value problem

$$\begin{aligned} &mx'' + kx = f(t), \qquad (0 \le t < \infty) \\ &x(0) = x_0, \quad x'(0) = x_0' \end{aligned} \tag{3.1}$$

to an integral equation, as follows. Integrate the differential equation from 0 to t twice. Using the initial conditions, show that those steps give

$$\begin{aligned} mx(t) - mx_0 - mx_0'\, t + k \textstyle\int_0^t \int_0^{t'} x(\tau)\, d\tau\, dt' \\ = \textstyle\int_0^t \int_0^{t'} f(\tau)\, d\tau\, dt'. \end{aligned} \tag{3.2}$$

Show that, by interchanging the order of integration, the double integrals can be reduced to single integrals, so that the integral equation (3.2) can be simplified to the form

$$\begin{aligned} mx(t) - mx_0 - mx_0'\, t + k \textstyle\int_0^t (t-\tau)\, x(\tau)\, d\tau \\ = \textstyle\int_0^t (t-\tau)\, f(\tau)\, d\tau. \end{aligned} \tag{3.3}$$

(b) Taking a Laplace transform of (3.3), obtain

$$X(s) = \frac{sx(0) + x'(0)}{s^2 + \omega^2} + \frac{F(s)}{m(s^2+\omega^2)}, \qquad \left(\omega = \sqrt{k/m}\right)$$

which is the same as equation (10).

4. Convert the initial-value problem

$$\begin{aligned} &mx'' + cx' + kx = f(t) \qquad (0 \le t < \infty) \\ &x(0) = x_0, \quad x'(0) = x_0' \end{aligned}$$

to an integral equation, analogous to (3.3) in Exercise 3. Then, solve that integral equation for $x(t)$ by using the Laplace transform.

5. (*Variable-coefficient equation*) Consider the problem

$$\begin{aligned} &tx'' + x' + tx = 0 \qquad (0 \le t < \infty) \\ &x(0) = 1, \quad x'(0) = 0, \end{aligned} \tag{5.1}$$

where our special interest lies in seeing whether or not we can solve (5.1) by the Laplace transform method even though the differential equation has nonconstant coefficients.

(a) Take the Laplace transform of the differential equation. Note that the transforms of $t\,x''(t)$ and $t\,x(t)$,

$$L\{t\,x''(t)\} = \int_0^\infty t\,x''\,e^{-st}\,dt,$$

$$L\{t\,x(t)\} = \int_0^\infty t\,x\,e^{-st}\,dt,$$

present a difficulty in that we cannot express them in terms of $X(s)$ the way we can express $L\{x'(t)\} = sX(s) - x(0)$ and $L\{x''(t)\} = s^2X(s) - sx(0) - x'(0)$. Nevertheless, these terms can be handled as follows. Observe that

$$\begin{aligned} L\{t\,x''(t)\} &= \int_0^\infty t\,x''\,e^{-st}\,dt = -\int_0^\infty \frac{d}{ds}\left(x''e^{-st}\right)dt \\ &= -\frac{d}{ds}\int_0^\infty x''e^{-st}\,dt \\ &= -\frac{d}{ds}\left[s^2X(s) - sx(0) - x'(0)\right] \\ &= -\frac{d}{ds}\left[s^2X(s) - s\right], \end{aligned} \tag{5.2}$$

if we assume that the unknown $x(t)$ is sufficiently well behaved for the third equality (where we have interchanged the order of two limit processes, the s differentiation and the t integration) to be justified. Handling the $L\{t\,x(t)\}$ term in the same way, show that application of the Laplace transform to (5.1) leads to the equation

$$(s^2+1)\frac{dX}{ds} + sX = 0 \tag{5.3}$$

on $X(s)$. Note that whereas the Laplace transform method reduces constant-coefficient differential equations to linear algebraic equations on $X(s)$, here the nonconstant coefficients result in the equation on $X(s)$ being itself a linear *differential* equation! However, it is a simple one. Solving (5.3), show that

$$X(s) = \frac{C}{\sqrt{s^2+1}}. \tag{5.4}$$

(b) From Appendix C, we find the inverse as $x(t) = CJ_0(t)$, where J_0 is the Bessel function of the first kind, of order zero. Appying the initial condition once again gives $x(0) = 1 = CJ_0(0) = C$, so $C = 1$, and the desired solution of (5.1) is $x(t) = J_0(t)$. Here, however, we ask you to proceed as though you don't know about Bessel functions. Specifically, re-express (5.4) as

$$X(s) = \frac{C}{s\sqrt{1+(1/s^2)}} = \frac{C}{s}\left(1 - \frac{1}{2}\frac{1}{s^2} + \cdots\right), \tag{5.5}$$

where the last equality amounts to the Taylor expansion of $\sqrt{1+r}$ in the quantity r, about $r = 0$, where $r = 1/s^2$. Carry that expansion further; invert the resulting series term by term (assuming that that step is valid), and thus show that

$$x(t) = C\left[1 - \frac{t^2}{2^2} + \frac{1}{(2!)^2}\frac{t^4}{2^4} - \frac{1}{(3!)^2}\frac{t^6}{2^6} + \cdots\right].$$

Setting $x(0) = 1$ gives $C = 1$, and the result is that we have obtained the solution in power series form. Of course, that power series is the Taylor series of the Bessel function $J_0(t)$. NOTE: Observe that rather than pulling an s out of the square root in (5.5), and then expanding $1/\sqrt{1+(1/s^2)}$ in powers of $1/s^2$, we could have expanded (5.4) directly in powers of s as $X(s) = C(1 - \frac{1}{2}s^2 + \cdots)$. However, positive powers of s are not invertible, so this form is of no use. [We will see, in Theorem 5.7.6, that to be invertible a transform must tend to zero as $s \to \infty$. Positive powers of s do not satisfy this condition, but negative powers do.] Also, observe that the degree to which nonconstant-coefficient differential equations are harder than constant-coefficient ones can be glimpsed from the fact that coefficients proportional to t cause the equation on $X(s)$ to be a first-order differential equation; coefficients proportional to t^2 will cause the equation on $X(s)$ to be a second-order differential equation, and so on.

6. It is found that the integral equation

$$C(T) = \int_0^\infty e^{-0.0744\nu^2/T^2}\rho(\nu)\,d\nu \tag{6.1}$$

is an approximate relation between the frequency spectrum $\rho(\nu)$ and the specific heat $C(T)$ of a crystal, where T is the temperature. Solve for $\rho(\nu)$ if

(a) $C(T) = T$ (b) $C(T) = Te^{-1/T}$

HINT: By a suitable change of variables, the integral can be made to be a Laplace transform.

7. We have seen that two crucial properties of the Laplace transform are its linearity and the property that $L\{f'(t)\} = s\overline{f}(s) - f(0)$; that is, the transform of the derivative is of the simple form $L\{f'(t)\} = a\overline{f}(s) + b$. With these properties in mind, consider the general integral transform

$$F(s) = \int_c^d K(t,s)f(t)\,dt \tag{7.1}$$

[equation (1) in Section 5.1] from a "design" point of view: how to choose the limits c, d and the kernel $K(t,s)$ to achieve these properties. Since $0 < t < \infty$, it is reasonable to choose $c = 0$ and $d = \infty$. Further, (7.1) automatically satisfies the linearity property $L\{\alpha u(t) + \beta v(t)\} = \alpha L\{u(t)\} + \beta L\{v(t)\}$ because the right side of (7.1) is an integral and integrals satisfy the property of linearity. Thus, we simply ask you for a logical derivation of the choice $K(t,s) = e^{-st}$ so that $L\{f'(t)\}$ is of the form $a\overline{f}(s) + b$.

5.5 Discontinuous Forcing Functions; Heaviside Step Function

Although we show in Section 5.2 that a given function has a Laplace transform if it is piecewise continuous on $0 \le t < A$ for every A and of exponential order as $t \to \infty$, we have thus far avoided functions with discontinuities. In applications, however, systems are often subjected to discontinuous forcing functions. For instance, a circuit might be subjected to an applied voltage that is held constant at 12 volts for a minute and then shut off (i.e., reduced to zero for all subsequent time). In this section we study systems with forcing functions that are discontinuous, although we still assume that they are piecewise continuous on $0 \le t < A$ for every A and of exponential order as $t \to \infty$, so that they are Laplace transformable.

We begin by defining the **Heaviside step function*** or **unit step function** (Fig. 1a),

$$\boxed{H(t) = \begin{cases} 0, & t < 0 \\ 1, & t \ge 0 \end{cases}} \tag{1}$$

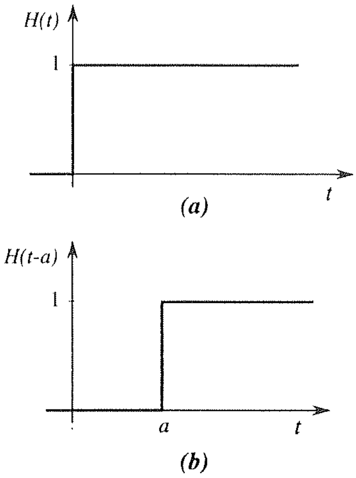


Figure 1. Unit step function.

which is a basic building block for our discussion. The value of $H(t)$ at $t = 0$ (i.e., at the jump discontinuity) is generally inconsequential in applications. We have chosen $H(0) = 1$ somewhat arbitrarily, and do not show the value of $H(t)$ at $t = 0$ in Fig. 1a to suggest that it is unimportant in this discussion.

Since $H(t)$ is a unit step at $t = 0$, $H(t - a)$ is a unit step shifted to $t = a$, as shown in Fig. 1b. In fact, the step function is useful in building up more complicated cases. We begin with the **rectangular pulse** shown in Fig. 2. Denoting that function as $P(t; a, b)$, we have

$$P(t; a, b) = H(t - a) - H(t - b). \tag{2}$$

More generally, observe that any piecewise continuous function

$$f(t) = \begin{cases} f_1(t), & 0 < t < t_1 \\ f_2(t), & t_1 < t < t_2 \\ \vdots & \vdots \\ f_n(t), & t_n < t < \infty \end{cases} \tag{3}$$

defined on $0 < t < \infty$ (which is the interval of interest in Laplace transform applications) can be given by the single expression

$$f(t) = f_1(t)\,P(t; 0, t_1) + \cdots + f_{n-1}(t)\,P(t; t_{n-1}, t_n) + f_n(t)\,H(t - t_n). \tag{4}$$

*Oliver Heaviside (1850–1925), initially a telegraph and telephone engineer, is best known for his contributions to vector field theory and to the development of a systematic Laplace transform methodology for the solution of differential equations. Note the spelling: Heaviside, not Heavyside.

In $0 < t < t_1$, for instance, each P function in (4) is zero except for the first, which equals unity in that interval; also, $H(t - t_n)$ is zero there, so (4) gives $f(t) = f_1(t)$. Similarly, in $t_1 < t < t_2, \dots$, and $t_{n-1} < t < t_n$. In $t_n < t < \infty$, each P function is zero and the $H(t - t_n)$ is unity, so (4) gives $f(t) = f_n(t)$ there.

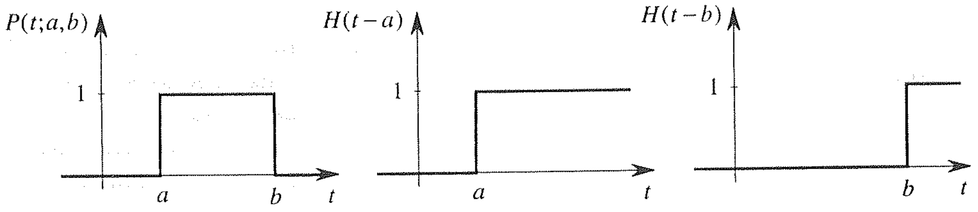


Figure 2. Rectangular pulse.

Note that (3) does not define $f(t)$ at the endpoints $0, t_1, \dots, t_n$. The Laplace transform of f will be the same no matter what those values are (assuming that they are finite) since the transform is an integral, an integral represents area, and there is no area under a finite number of points. Thus, those values will be inconsequential; hence we don't even specify them in (3).

EXAMPLE 1. The function

$$f(t) = \begin{cases} 2 + t^2, & 0 < t < 2 \\ 6, & 2 < t < 3 \\ 2/(2t - 5) & 3 < t < \infty \end{cases} \tag{5}$$

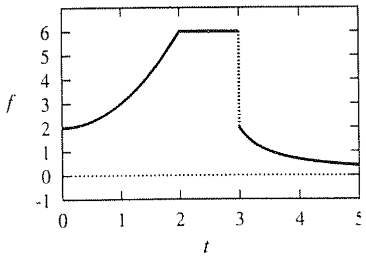


Figure 3. $f(t)$ of Example 1.

shown in Fig. 3, can be expressed, according to (4), as

$$f(t) = (2 + t^2)\,[H(t) - H(t - 2)] + 6\,[H(t - 2) - H(t - 3)] + \frac{2}{2t - 5}\,H(t - 3). \tag{6}$$

Actually, since the interval is $0 < t < \infty$ we cannot distinguish between $H(t)$ and unity, so we could replace the $H(t)$ in the first term by 1. ▮

EXAMPLE 2. *Ramp function.* The function

$$f(t) = \begin{cases} 0, & 0 < t < a \\ t - a & a < t < \infty \end{cases} \tag{7}$$

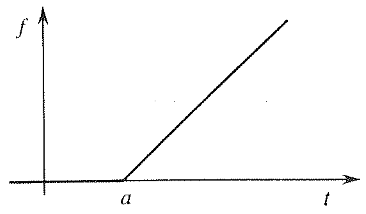


Figure 4. The ramp function of Example 2.

shown in Fig. 4 is called a **ramp** function and, according to (4), it can be expressed as $f(t) = (t - a)H(t - a)$. ▮

Before considering applications, observe that

$$L\{H(t - a)\} = \int_0^\infty H(t - a)\, e^{-st}\, dt = \int_a^\infty e^{-st}\, dt = \frac{e^{-as}}{s},$$

so the Laplace transform of $H(t-a)$ is

$$\boxed{L\{H(t-a)\} = \frac{e^{-as}}{s}.} \tag{8}$$

Also important to us is the result

$$\boxed{L\{H(t-a)f(t-a)\} = e^{-as}F(s)} \tag{9a}$$

or, equivalently,

$$\boxed{L^{-1}\left\{e^{-as}F(s)\right\} = H(t-a)f(t-a)} \tag{9b}$$

for any (Laplace-transformable) function $f(t)$. Proof is as follows:

$$\begin{aligned} L\{H(t-a)f(t-a)\} &= \int_0^\infty H(t-a)\,f(t-a)\,e^{-st}\,dt \\ &= \int_a^\infty f(t-a)\,e^{-st}\,dt = \int_0^\infty f(\tau)\,e^{-s(\tau+a)}\,d\tau \\ &= e^{-as}\int_0^\infty f(\tau)\,e^{-s\tau}\,d\tau = e^{-as}F(s), \end{aligned} \tag{10}$$

where the third equality follows the change of variables $t-a=\tau$. In words, $H(t-a)f(t-a)$ is the function $f(t)$ delayed by a time interval a, as illustrated in Fig. 5 for the function $f(t)=\sin t$.

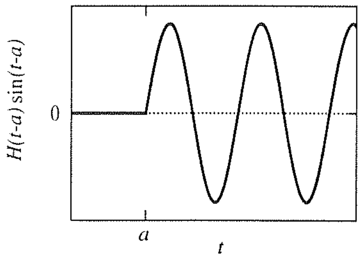


Figure 5. Delay significance of $H(t-a)f(t-a)$.

EXAMPLE 3. *LC Circuit.* We saw in Section 2.3 that the differential equation governing the charge $Q(t)$ on the capacitor in the circuit shown (Fig. 6) is $LQ''+RQ'+(1/C)Q=E(t)$. Let $R=0$ and let $E(t)$ be the rectangular pulse shown in Fig. 2, of magnitude E_0, and with $a=2$ and $b=5$, for definiteness. Thus, $E(t)=E_0\,[H(t-2)-H(t-5)]$. If $Q(0)=Q_0$ and $Q'(0)=0$, then we have the initial-value problem

$$LQ''+\frac{1}{C}Q = E_0\,[H(t-2)-H(t-5)], \tag{11a}$$

$$Q(0)=Q_0, \qquad Q'(0)=0 \tag{11b}$$

on $Q(t)$. [Since the current $i(t)$ is dQ/dt, $Q'(0)=0$ means that $i(0)=0$ so we can think of a switch being open until time $t=0$, and closed at that instant.] We wish to solve (11) for $Q(t)$. Be careful: we will need to distinguish the inductance L from the Laplace transform L by the context.

Taking a Laplace transform of (11a), and using (11b) and (8), gives

$$L\left(s^2\overline{Q}(s)-sQ_0\right)+\frac{1}{C}\overline{Q}(s) = E_0\left(\frac{e^{-2s}}{s}-\frac{e^{-5s}}{s}\right) \tag{12}$$

so

$$\overline{Q}(s) = \frac{Q_0 s}{s^2+\omega^2}+\frac{E_0}{L}\frac{1}{s(s^2+\omega^2)}\left(e^{-2s}-e^{-5s}\right), \tag{13}$$

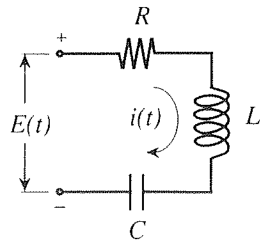


Figure 6. *RLC* circuit.

where $\omega = 1/\sqrt{LC}$. [Generally, we use the notation $L\{f(t)\} = F(s)$, but in $Q(t)$ the Q is already capitalized, so we use $L\{Q(t)\} = \overline{Q}(s)$ instead.] To invert (13), we begin with

$$L^{-1}\left\{\frac{1}{s(s^2+\omega^2)}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s^2+\omega^2}\right\} = 1 * \frac{\sin\omega t}{\omega}$$
$$= \int_0^t \frac{\sin\omega\tau}{\omega}\,d\tau = \frac{1-\cos\omega t}{\omega^2}. \tag{14}$$

Then, using (14) and (9b) and $L^{-1}\left\{s/\left(s^2+\omega^2\right)\right\} = \cos\omega t$ from Appendix C, we have

$$Q(t) = Q_0\cos\omega t + E_0C\left\{H(t-2)\left[1-\cos\omega(t-2)\right]\right.$$
$$\left.-H(t-5)\left[1-\cos\omega(t-5)\right]\right\}, \tag{15}$$

which is shown in Fig. 7 for the representative case where $Q_0 = E_0 = L = C = 1$.

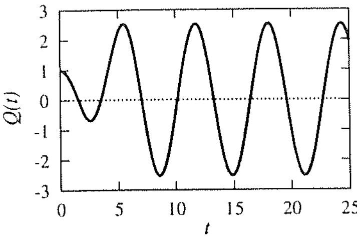


Figure 7. $Q(t)$ given by (15).

COMMENT 1. Most striking is the way the use of the Heaviside notation and the Laplace transform have enabled us to solve for $Q(t)$ on the entire t domain $(0 < t < \infty)$. In contrast, if we rely on the methods of Chapter 3 we need to break (11) into three separate problems:

$$\begin{array}{llll} 0 \le t \le 2: & LQ'' + (1/C)Q = 0, & Q(0) = Q_0, & Q'(0) = 0 \\ 2 \le t \le 5: & LQ'' + (1/C)Q = E_0, & Q(2) = ?, & Q'(2) = ? \\ 5 \le t < \infty: & LQ'' + (1/C)Q = 0, & Q(5) = ?, & Q'(5) = ? \end{array} \tag{16a,b,c}$$

First, we solve (16a) for $Q(t)$ on $0 \le t \le 2$. The final values from that solution, $Q(2)$ and $Q'(2)$, then serve as the initial conditions for the next problem, (16b). Then we solve for $Q(t)$ on $2 \le t \le 5$ and use the final values from that solution, $Q(5)$ and $Q'(5)$, as the initial conditions for the next problem, (16c). Clearly, this approach is more tedious than the Laplace transform approach that led to (15).

COMMENT 2. A fundamental question comes to mind: Does the discontinuous nature of the input $E(t)$ result in the output $Q(t)$ being discontinuous as well? We can see from the graph in Fig. 7 that the answer is no. The continuity of $Q(t)$ may be surprising from the solution form (15), because of the presence of the two Heaviside functions in (15). However, the jump discontinuity implied by the $H(t-2)$ is eliminated by its $1-\cos\omega(t-2)$ factor since the latter vanishes at $t = 2$. Similarly, the jump discontinuity that is implied by the $H(t-5)$ is eliminated by its $1-\cos\omega(t-5)$ factor since the latter vanishes at $t = 5$. ▮

To better understand how a discontinuous input can produce a continuous output, consider the following simplified situation, the equation

$$Q''(t) = H(t-a) \tag{17}$$

with discontinuous right-hand side. Integrating (17) gives

$$Q'(t) = (t-a)H(t-a) + A, \tag{18}$$

because the derivative of the right-hand side is indeed $H(t-a)$, as can be seen from Fig. 4. Integrating again, we obtain

$$Q(t) = \frac{(t-a)^2}{2}H(t-a) + At + B, \tag{19}$$

and these results are shown in Fig. 8 (for the case where $A = B = 0$, say). The idea is that a differential equation is solved by a process which, essentially, involves integration, and integration is a smoothing process! For observe that whereas $Q''(t) = H(t-a)$ is discontinuous at $t = a$, $Q'(t) = (t-a)H(t-a)$ is continuous but with a "kink," and $Q(t) = (t-a)^2 H(t-a)/2$ is continuous and smooth (differentiable) as well.

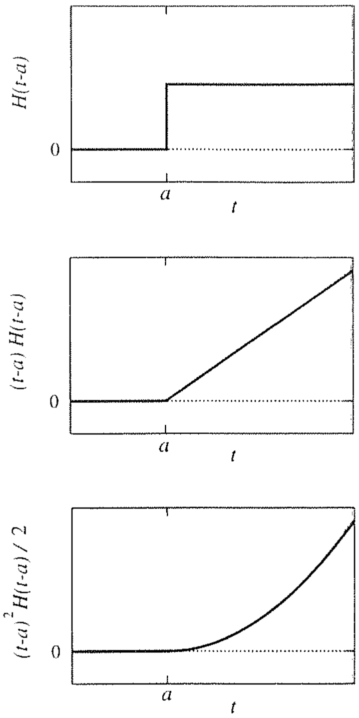


Figure 8. The smoothing effect of integration.

EXAMPLE 4. *RC Circuit.* In Example 3 we took $R = 0$ in the circuit shown in Fig. 6, and considered the resulting LC circuit. Here, let us take $L = 0$ instead, and consider the resulting RC circuit, governed by the first-order equation $RQ' + (1/C)Q = E(t)$. Further, let $Q(0) = 0$ and let $E(t) = 50t$ on $0 < t < 2$ and $E(t) = 40$ on $2 < t < \infty$ (sketch it). According to (4) then, $E(t) = 50t\,[1 - H(t-2)] + 40H(t-2)$. Let $R = C = 1$, for simplicity. Then the initial-value problem on $Q(t)$ is

$$Q' + Q = 50t + (40 - 50t)H(t-2), \tag{20a}$$

$$Q(0) = 0. \tag{20b}$$

Laplace transforming (20a),

$$s\overline{Q}(s) + \overline{Q}(s) = \frac{50}{s^2} + L\left\{(40-50t)H(t-2)\right\}, \tag{21}$$

where

$$\begin{aligned} L\left\{(40-50t)H(t-2)\right\} &= L\left\{[-60 - 50(t-2)]\,H(t-2)\right\} \\ &= -60L\left\{H(t-2)\right\} - 50L\left\{(t-2)H(t-2)\right\} \\ &= -60\frac{e^{-2s}}{s} - 50e^{-2s}L\{t\} = -60\frac{e^{-2s}}{s} - 50\frac{e^{-2s}}{s^2}. \end{aligned} \tag{22}$$

Putting (22) into (21) and solving for $\overline{Q}(s)$ gives

$$\overline{Q}(s) = \frac{50}{s^2(s+1)} - \frac{60}{s(s+1)}e^{-2s} - \frac{50}{s^2(s+1)}e^{-2s}, \tag{23}$$

which we now need to invert. Taking one term at a time,

$$L^{-1}\left\{\frac{1}{s(s+1)}\right\} = L^{-1}\left\{\frac{1}{s}\right\} * L^{-1}\left\{\frac{1}{s+1}\right\} = 1 * e^{-t} = \int_0^t e^{-\tau}\,d\tau = 1 - e^{-t}, \tag{24a}$$

$$\begin{aligned} L^{-1}\left\{\frac{1}{s^2(s+1)}\right\} &= L^{-1}\left\{\frac{1}{s^2}\right\} * L^{-1}\left\{\frac{1}{s+1}\right\} = t * e^{-t} \\ &= \int_0^t (t-\tau)\,e^{-\tau}\,d\tau = t - 1 + e^{-t}, \end{aligned} \tag{24b}$$

$$L^{-1}\left\{\frac{1}{s(s+1)}e^{-2s}\right\}=\left(1-e^{-(t-2)}\right)H(t-2), \quad \text{from (9b), (24a)} \tag{24c}$$

$$L^{-1}\left\{\frac{1}{s^2(s+1)}e^{-2s}\right\}=\left[(t-2)-1+e^{-(t-2)}\right]H(t-2), \quad \text{from (9b), (24b)} \tag{24d}$$

so

$$\begin{aligned} Q(t) &= 50\left(t-1+e^{-t}\right)-60\left(1-e^{-(t-2)}\right)H(t-2) \\ &\quad -50\left[(t-2)-1+e^{-(t-2)}\right]H(t-2) \\ &= 50\left(t-1+e^{-t}\right)+\left(90-50t+10e^{2-t}\right)H(t-2) \end{aligned}$$

is the desired particular solution of (20). ▮

Computer software. The *Maple* name for $H(t)$ is Heaviside(t).

Closure. We introduce the Heaviside step function $H(t)$ and show, in equation (4), how to use it in representing piecewise-continuous functions. A key result involving the Heaviside function is given by (9). Finally, we show how convenient the Heaviside notation is, together with the Laplace transform, in solving differential equations with piecewise continuous forcing functions.

EXERCISES 5.5

1. Use (4) to give a single expression for $f(t)$, and give a labeled sketch of its graph, as well. From that expression, evaluate the transform $F(s)$ of $f(t)$.

(a) $f(t)=t$ on $0<t<2$, $4-t$ on $2<t<4$, and 0 on $t>4$
(b) $f(t)=e^{-t}$ on $0<t<1$, 0 on $t>1$
(c) $f(t)=2$ on $0<t<5$, -3 on $5<t<7$, 1 on $t>7$
(d) $f(t)=t^3-t$ on $0<t<1$, -6 on $t>1$
(e) $f(t)=2-t$ on $0<t<2$, $2t-6$ on $2<t<5$, t on $t>5$
(f) $f(t)=t^3$ on $0<t<10$, $3t^2-2t$ on $10<t<20$, $5t$ on $t>20$
(g) $f(t)=\sin t$ on $0<t<5\pi$, 0 on $t>5\pi$
(h) $f(t)=\cos t$ on $0<t<\pi$, -1 on $t>\pi$

2. Draw a labeled sketch of the graph of each function.

(a) $H(t-1)e^{t-1}$
(b) $H(t-2\pi)\cos(t-2\pi)$
(c) $(1+t)H(t-2)$
(d) $(2+t)\left[H(t-2)-H(t-3)\right]$
(e) $t\left[H(t-1)-H(t-2)+H(t-3)\right]$
(f) $t^2\left[2H(t-1)-H(t-3)-H(t-4)\right]$
(g) $\left[H(t-\pi/2)-H(t-\pi)\right]\sin t$
(h) $1+H(t-1)+H(t-2)+H(t-3)+H(t-4)$
(i) $H(t-3)[H(t-2)-H(t-1)]$

3. Evaluate in terms of Heaviside functions. You may use these results for the definite and indefinite integrals of the Heaviside function:

$$\boxed{\int_0^t H(\tau)\,d\tau=\begin{cases}0, & t<0\\ t, & t>0\end{cases}=tH(t).} \tag{3.1}$$

and

$$\boxed{\int^t H(\tau)\,d\tau=tH(t)+\text{constant.}} \tag{3.2}$$

(a) $\int_{-5}^t\left[H(\tau)-2\right]d\tau$
(b) $\int_0^t \tau\,H(\tau-2)\,d\tau$
(c) $\int_0^t\left[1-H(\tau-5)\right]d\tau$
(d) $\int_0^t\left[H(\tau-a)-H(\tau-b)\right]d\tau \qquad (b>a)$
(e) $\int_t^\infty\left[H(\tau-2)-H(\tau-3)\right]d\tau$
(f) $\int_t^{2t} H(\tau-1)\,d\tau$

(g) $\int_0^5 H(\tau - t)\,d\tau$
(h) $t * H(t-1)$
(i) $\sin t * [H(t-1) - H(t-2)]$
(j) $e^{-t} * H(t-5)$
(k) $1 * H(t-1)$

4.(a)–(k) Evaluate the integral in the corresponding part of Exercise 3 using computer software such as the *Maple* int command.

5. Solve $x' - x = f(t)$, where $x(0) = 0$, by the methods of this section, where $f(t)$ is:

(a) $H(t-1)$
(b) $e^{-t}H(t-3)$
(c) t on $0 < t < 2$, 2 on $t > 2$
(d) 0 on $0 < t < 5$, 10 on $5 < t < 7$, 0 on $t > 7$
(e) 0 for $t \neq 5$, 100 for $t = 5$
(f) $1 - e^{-t}$ on $0 < t < 6$, 0 on $t > 6$
(g) 0 on $0 < t < 1$, 1 on $1 < t < 2$, 2 on $2 < t < 3$, 1 on $3 < t < 4$, 0 on $t > 4$
(h) t on $0 < t < 1$, $2 - t$ on $1 < t < 2$, 0 on $t > 2$
(i) e^{-t} on $0 < t < 1$, e^{1-t} on $t > 1$
(j) 20 on $0 < t < 1$, 10 on $1 < t < 2$, 0 on $t > 2$

6.(a) – (j) Same as Exercise 5, but using computer software such as the *Maple* dsolve command.

7.(a) – (j) Same as Exercise 5, but for $x'' - x = f(t)$, $x(0) = x'(0) = 0$.

5.6 Impulsive Forcing Functions; Dirac Impulse Function (Optional)

Besides forcing functions that are discontinuous, we are interested in ones that are impulsive – that is, sharply focused in time. For instance, consider the forced mechanical oscillator governed by the differential equation

$$mx'' + cx' + kx = f(t), \tag{1}$$

where $f(t)$ is the force applied to the mass. If the force is due to a hammer blow, for instance, initiated at time $t = 0$, then we expect $f(t)$ to be somewhat as sketched in Fig. 1a. However, we do not know the functional form of $f(t)$ corresponding to such an event as a hammer blow, so the problem that we pose is how to proceed with the solution of (1) without knowing f. Of course we can solve (1) in terms of f, but eventually we need to know f to find the response $x(t)$.

In working with impulsive forces one normally tries to avoid dealing with the detailed shape of f and tries to limit one's concern to a global quantity known as the **impulse** I of f, the area under its graph. The idea is that if ϵ really is small, then the response $x(t)$, while sensitive to I, should be rather insensitive to the detailed shape of f. That is, if we vary the shape of f but keep its area the same, then we expect little change in the response $x(t)$. This idea suggests that we replace the unknown f by a simple rectangular pulse having the correct impulse as shown in Fig. 1b: $f(t) = I/\epsilon$ for $0 \leq t \leq \epsilon$, and 0 for $t > \epsilon$. With f thus simplified we can proceed to solve for the response $x(t)$. But even so, the solution still depends upon ϵ, and the latter is probably not known, just as the actual shape of f is not known. Thus, we adopt one more idealization: we suppose that since ϵ is very small, we might as well take the limit of the solution as $\epsilon \to 0$, to eliminate ϵ.

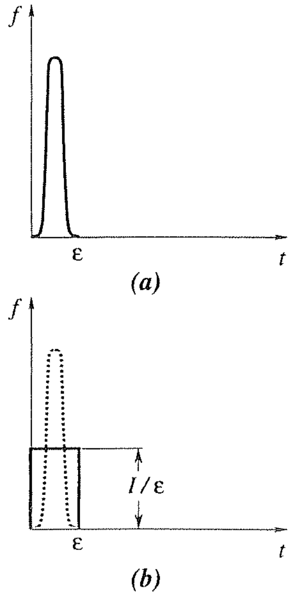


Figure 1. Impulsive force at $t = 0$.

Let us denote such a rectangular pulse having a unit impulse ($I = 1$) as $D(t;\epsilon)$:

$$D(t;\epsilon) = \begin{cases} 1/\epsilon, & 0 \le t \le \epsilon \\ 0, & t > \epsilon, \end{cases} \tag{2}$$

where we use D (after the physicist *P. A. M. Dirac*, who developed the idea of impulsive forces in 1929). As $\epsilon \to 0$, D becomes taller and narrower as shown in Fig. 2, in such a way as to maintain its unit area. Of course, the limit

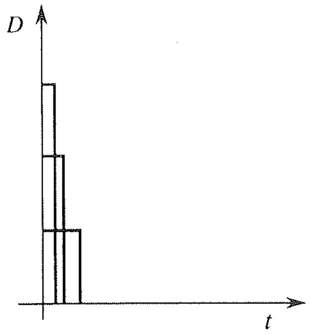


Figure 2. Letting $\epsilon \to 0$ in (2).

$$\lim_{\epsilon\to 0} D(t;\epsilon) = \begin{cases} \infty, & t = 0 \\ 0, & t > 0 \end{cases} \tag{3}$$

does not exist, because ∞ is not an acceptable value, but Dirac showed that it is nevertheless useful to think of that limiting case as representing an idealized point unit impulse focused at $t = 0$.

To explain, we first prove that

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = g(0) \tag{4}$$

for any function g that is continuous at the origin. To begin our proof, write

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \int_0^\epsilon g(\tau)\, \frac{1}{\epsilon}\, d\tau. \tag{5}$$

Suppose that g is continuous on $0 \le \tau \le b$ for some positive b. We can assume that $\epsilon < b$ because we are letting $\epsilon \to 0$. Thus, g is continuous on the integration interval $0 \le \tau \le \epsilon$, so the mean value theorem of the integral calculus tells us that there is a point τ_1 in $[0,\epsilon]$ (i.e., the closed interval $0 \le \tau \le \epsilon$) such that $\int_0^\epsilon g(\tau)\, d\tau = g(\tau_1)\epsilon$. Thus, (5) gives

$$\lim_{\epsilon\to 0} \int_0^\infty g(\tau)\, D(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \frac{1}{\epsilon} g(\tau_1)\epsilon = \lim_{\epsilon\to 0} g(\tau_1) = g(0), \tag{6}$$

where the last equality holds since τ_1 is in the interval $[0,\epsilon]$, and ϵ is going to zero. Finally, since b is arbitrarily small, we only need the continuity of g *at* $\tau = 0$. This completes our proof of (4).

For brevity, it is customary to dispense with calling attention to the ϵ limit and to express (4) as

$$\int_0^\infty g(\tau)\, \delta(\tau)\, d\tau = g(0), \tag{7}$$

where $\delta(\tau)$ is known as the **Dirac delta function**, or **unit impulse function**. We can think of $\delta(\tau)$ as being zero everywhere except at the origin and infinite at the origin, in such a way as to have unit area, but it must be noted that that definition is not satisfactory within the framework of ordinary function theory. To create a legitimate place for the delta function, one needs to extend the concept of function.

That was done by *L. Schwartz*, and the result is known as the *theory of distributions*, but that theory is well beyond our present scope.

Let us illustrate the application of the delta function with an example.

EXAMPLE 1. Consider (1), with $m = k = 1$ and $c = 0$; let $f(t)$ correspond to a hammer blow as sketched in Fig. 1a, and let $x(0) = x'(0) = 0$, so that before the blow the mass is at rest. The solution of the problem

$$x'' + x = f(t), \qquad x(0) = x'(0) = 0 \tag{8}$$

is found, for instance, by using the Laplace transform, to be

$$x(t) = \int_0^t \sin(t-\tau)\, f(\tau)\, d\tau. \tag{9}$$

As outlined above, the idea is to replace $f(\tau)$ by a rectangular pulse $ID(\tau;\epsilon)$ having the same area I as $f(\tau)$ and then to take the limit as $\epsilon \to 0$:

$$\begin{aligned} x(t) &= \lim_{\epsilon\to 0} \int_0^t \sin(t-\tau)\, ID(\tau;\epsilon)\, d\tau = \lim_{\epsilon\to 0} \int_0^\epsilon \sin(t-\tau)\, \frac{I}{\epsilon}\, d\tau \\ &= \lim_{\epsilon\to 0} I \frac{\cos(t-\epsilon) - \cos t}{\epsilon} = I \sin t, \end{aligned} \tag{10}$$

where the last equality follows from l'Hôpital's rule.

Alternatively and more simply, let $f(\tau) = I\delta(\tau)$ in (9), where the scale factor I is needed since the delta function is a unit impulse whereas we want the impulse to be I. Then property (7) of the delta function gives

$$x(t) = \int_0^t \sin(t-\tau)\, I\delta(\tau)\, d\tau = I \sin(t-\tau)\Big|_{\tau=0} = I \sin t, \tag{11}$$

as obtained previously in (10). You may be concerned that we have applied (7) even though the upper integration limits in (7) and (9) are not the same. However, in (5) we see that the ∞ was immediately changed to ϵ, and then we let ϵ tend to zero. Thus, (7) holds for any positive upper limit; we used ∞ just for definiteness. ▮

Let us review the idea. Since, generally, we know neither the exact shape nor the duration of an impulsive forcing function f, we do two things to solve for the response. We replace f by an equivalent rectangular pulse (i.e., having the same impulse, or area, as f), solve for $x(t)$, and then we let the width of the pulse, ϵ, tend to zero. Equivalently and more simply, we take f to be a Dirac delta function and evaluate the resulting integral using the fundamental property (7) of the delta function. The latter procedure is more efficient because one no longer needs to take the limit of the integral as $\epsilon \to 0$; the limit was already carried out, once and for all, in our derivation of (4).

Since $\delta(t)$ is focused at $t = 0$, it follows that $\delta(t-a)$ is focused at $t = a$, and (7) generalizes to

$$\int_0^\infty g(\tau)\, \delta(\tau - a)\, d\tau = g(a). \tag{12}$$

Here we continue to use the $0, \infty$ limits, but it should be understood that the result is $g(a)$ for any limits A, B (with $B > A$) such that the point of action of the delta function is contained within the interval of integration. If the point $t = a$ falls outside the interval, then the integral is zero. Thus, for reference, we give the following more complete result:*

$$\boxed{\int_A^B g(\tau)\,\delta(\tau - a)\,d\tau = \begin{cases} g(a), & A \le a < B \\ 0, & a < A \quad \text{or} \quad a \ge B. \end{cases}} \tag{13}$$

EXAMPLE 2. *RC Circuit.* Recall from Section 5.5 that the charge $Q(t)$ on the capacitor of the RC circuit is governed by the differential equation $RQ' + (1/C)Q = E(t)$. Let $E(t)$ be an impulsive voltage, with impulse I acting at $t = T$, and let $Q(0) = Q_0$. We wish to solve for $Q(t)$. Expressing $E(t) = I\delta(t - T)$, the initial-value problem is

$$Q' + \kappa Q = I\delta(t - T), \qquad Q(0) = Q_0, \tag{14}$$

where $\kappa = 1/(RC)$. Taking the Laplace transform of (14) gives $s\overline{Q} - Q_0 + \kappa\overline{Q} = IL\{\delta(t - T)\}$ so

$$\overline{Q} = \frac{Q_0}{s + \kappa} + I\frac{1}{s + \kappa}L\{\delta(t - T)\}, \tag{15}$$

and

$$\begin{aligned} Q(t) &= Q_0 e^{-\kappa t} + Ie^{-\kappa t} * \delta(t - T) \\ &= Q_0 e^{-\kappa t} + I\int_0^t e^{-\kappa(t-\tau)}\delta(\tau - T)\,d\tau \\ &= Q_0 e^{-\kappa t} + \begin{cases} 0, & t < T \\ Ie^{-\kappa(t-T)}, & t > T \end{cases} \\ &= Q_0 e^{-\kappa t} + IH(t - T)e^{-\kappa(t-T)}, \end{aligned} \tag{16}$$

where the third equality follows from (13). ■

Observe that we do not need to know the transform of the delta function in Example 2; we merely call its transform $L\{\delta(t - T)\}$, and inversion by the convolution theorem gives us back the $\delta(t - T)$ that we started with. Nonetheless, for reference, let us work out its Laplace transform. According to (12),

$$L\{\delta(t - a)\} = \int_0^\infty \delta(t - a)\,e^{-st}\,dt = e^{-st}\Big|_{t=a} = e^{-as}. \tag{17}$$

*Following (12), we state that the result is $g(a)$ if the delta function acts within the integration interval. How then do we interpret the integral when a is at an endpoint (A or B)? We've met that case in equation (7). Since the $D(\tau; \epsilon)$ sequence (Fig. 2) is defined on $[0, \epsilon]$, the delta function acts essentially to the right of $\tau = 0$, hence within the interval of integration, and the result of the integration is $g(0)$. To be consistent, let us suppose that the D sequence is always to the right of the point $\tau = a$. Then the integral in (13) will be $g(a)$ if $A \le a < B$ and 0 if $a < A$ or $a \ge B$.

In particular, $L\{\delta(t)\} = 1$.

Since this section is about the Laplace transform, the independent variable has been the time t, so the delta function has represented actions that are focused in time. But the argument of the delta function need not be time. For instance, if $w(x)$ is the load distribution on a beam (Fig. 3a), in pounds per unit length, then $\delta(x-a)$ represents a point unit load (i.e., one pound) at $x = a$ (Fig. 3b).

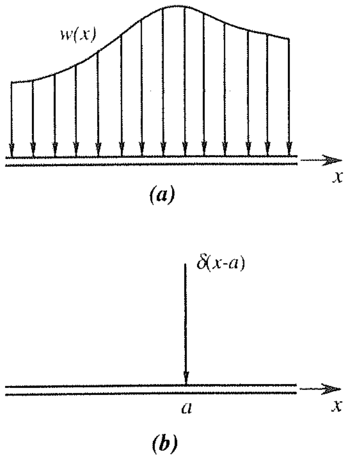


Figure 3. Load distribution on a beam.

Let us close this discussion with a comment on the delta function idealization from a modeling point of view. Consider a metal plate, extending over $-\infty < x < \infty$ and $0 < y < \infty$, loaded by pressing a metal coin against it, at the origin, with a force P (Fig. 4a). If one is to determine (from the theory of elasticity) the stress distribution within the plate, one needs to know the load distribution $w(x)$ along the edge of the plate (namely, the x axis). Because the coin will flatten slightly, at the point of contact, the load $w(x)$ will be distributed over a short interval, say from $x = -\epsilon$ to $x = \epsilon$. However, the function $w(x)$ is not known a priori and its determination is part of the problem. Whether one needs to determine the exact $w(x)$ distribution or if it suffices to represent it simply as an idealized point force of magnitude P, $w(x) = P\delta(x)$, depends upon whether one is interested in the "near field" or the "far field." By the near field we mean that part of the plate within several ϵ lengths of the point of the load application – for instance, within the dashed semicircle shown in Fig. 4b. The far field is the region beyond. If we are concerned only with the far field, then it should suffice to use

$$w(x) = P\delta(x), \tag{18}$$

but if concerned with the near field then the approximation (18) will lead to large errors. A ball bearing manufacturer, for instance, is primarily interested with the near field induced by a loaded ball bearing due to concern regarding wear and surface damage. Within the theory of elasticity, the insensitivity of the far field to the detailed shape of $w(x)$ [given that the area under the $w(x)$ graph is held fixed] is an example of *Saint Venant's principle*.

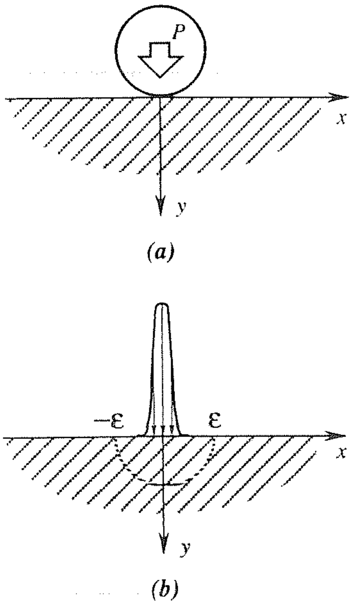


Figure 4. Delta function idealization.

Computer software. The *Maple* name for $\delta(t)$ is Dirac(t).

Closure. We introduce the delta function out of a need to deal effectively with impulsive forcing functions, functions that are highly focused in time or space. Often we know neither the precise form of such a function nor the precise interval of application. If that interval is short enough one can model the force as an idealized point force, represented mathematically as a delta function $\delta(t)$. One is not so much interested in the numerical values of $\delta(t)$ [indeed, one says that $\delta(0) = \infty$] as in the effect of integration upon a delta function, and that effect is expressed by (13), which we regard as the most important formula in this section.

EXERCISES 5.6

1. Solve for $x(t)$, on $0 \le t < \infty$.

(a) $x'' - x = \delta(t-2)$; $\quad x(0) = x'(0) = 0$
(b) $x'' - 4x = 6\delta(t-1)$; $\quad x(0) = 0,\ x'(0) = -3$
(c) $x'' - 3x' + 2x = 2 + \delta(t-5)$; $\quad x(0) = x'(0) = 0$
(d) $x'' + x' = 1 + \delta(t-2)$; $\quad x(0) = 0,\ x'(0) = 3$
(e) $x'' + 2x' + x = 10\delta(t-5)$; $\quad x(0) = x'(0) = 0$
(f) $2x'' - x' = \delta(t-1) - \delta(t-2)$; $\quad x(0) = x'(0) = 0$
(g) $x'' - 3x' + 2x = 100\delta(t-3)$; $\quad x(0) = 4,\ x'(0) = 0$
(h) $x''' = 2\delta(t-5)$; $\quad x(0) = x'(0) = x''(0) = 0$
(i) $x''' + 3x'' + 2x' = \delta(t-5)$; $\quad x(0) = x'(0) = x''(0) = 0$
(j) $x'''' - 4x'' = 3\delta(t-1)$; $\quad x(0) = x'(0) = x''(0) = 0$, $x'''(0) = 1$
(k) $x'''' - 5x'' + 4x = 6\delta(t-2)$; $\quad x(0) = x'(0) = x''(0) = x'''(0) = 0$
(l) $x'''' - x = \delta(t-1)$; $\quad x(0) = x'(0) = x''(0) = x'''(0) = 0$

2. Show that the delta function has these properties, where κ is a nonzero real constant, and the function $f(t)$ is continuous at the origin. NOTE: Recall that the delta function is defined by its integral behavior. Thus, by an equation such as $\delta(-t) = \delta(t)$ we mean that

$$\int_{-\infty}^{\infty} g(t)\,\delta(-t)\,dt = \int_{-\infty}^{\infty} g(t)\,\delta(t)\,dt \tag{2.1}$$

for every function $g(t)$ that is continuous at the origin. The right side of (2.1) is $g(0)$, so to show that $\delta(-t) = \delta(t)$, in part (a), you need to verify that the left side of (2.1) is $g(0)$ too.

(a)
$$\delta(-t) = \delta(t) \tag{2.2}$$

(b)
$$\delta(\kappa t) = \frac{1}{|\kappa|}\,\delta(t) \quad (\kappa \neq 0) \tag{2.3}$$

(c)
$$f(t)\delta(t) = \begin{cases} f(0)\delta(t), & f(0) \neq 0 \\ 0, & f(0) = 0 \end{cases} \tag{2.4}$$

For instance, $(3t+2)\delta(t) = 2\delta(t)$, $(\sin t)\delta(t) = 0$, $(3t+2)\delta(t-1) = 5\delta(t-1)$, and $(t^2+t-2)\delta(t-1) = 0$. Formally, the first part of (2.4) makes sense as follows: $\delta(t)$ is nonzero only at $t = 0$, so there is no difference between $f(t)\delta(t)$ and $f(0)\delta(t)$.

(d)
$$\boxed{\int_{-\infty}^{t} \delta(\tau)\,d\tau = H(t)} \tag{2.5}$$

3. The result (2.5), above, reveals the close relation between the delta and Heaviside functions. Alternatively, we can write that relation as

$$\boxed{H'(t) = \delta(t).} \tag{3.1}$$

The latter follows from (2.5) only in a formal sense, but is quite useful, along with (2.2)–(2.5). For instance, suppose we wish to verify that $x(t) = H(t-1)\sin(t-1)$ satisfies the initial-value problem $x'' + x = \delta(t-1)$; $\quad x(0) = x'(0) = 0$. Differentiating $x(t)$ gives

$$\begin{aligned} x'(t) &= H'(t-1)\sin(t-1) + H(t-1)\cos(t-1) \\ &= \delta(t-1)\sin(t-1) + H(t-1)\cos(t-1) \\ &= 0 + H(t-1)\cos(t-1), \end{aligned} \tag{3.2}$$

and

$$\begin{aligned} x''(t) &= H'(t-1)\cos(t-1) - H(t-1)\sin(t-1) \\ &= \delta(t-1)\cos(t-1) - H(t-1)\sin(t-1) \\ &= \delta(t-1) - H(t-1)\sin(t-1) \end{aligned} \tag{3.3}$$

so $x'' + x$ does give $\delta(t-1)$. In the second equality in (3.2) we used (3.1), and in the third we used (2.4): $\delta(t-1)\sin(t-1) = \delta(t-1)\sin 0 = 0$. In the second equality in (3.3) we used (3.1), and in the third we used (2.4): $\delta(t-1)\cos(t-1) = \delta(t-1)\cos 0 = \delta(t-1)$. Further, we see that $x(0) = 0$ and, from (3.2), that $x'(0) = 0$. Here is the problem: In the same manner as above, verify the following solutions that are given in the Answers to the Selected Exercises.

(a) exercise 1(a)
(b) exercise 1(d)
(c) exercise 1(g)

5.7 Additional Properties

In Section 5.3 we establish the linearity of the transform and its inverse, the transform of the derivative $f'(t)$, and the Laplace convolution theorem, results that we deem essential in applying the Laplace transform to the solution of differential equations. In this final section of Chapter 5 we present several additional useful properties of the Laplace transform.

THEOREM 5.7.1 *s-Shift*
If $L\left\{f(t)\right\} = F(s)$ exists for $s > s_0$, then for any real constant a,

$$L\left\{e^{-at}f(t)\right\} = F(s+a) \tag{1}$$

for $s + a > s_0$ or, equivalently,

$$L^{-1}\left\{F(s+a)\right\} = e^{-at}f(t). \tag{2}$$

Proof:

$$\begin{aligned} L\left\{e^{-at}f(t)\right\} &= \int_0^\infty e^{-at}f(t)\,e^{-st}\,dt \\ &= \int_0^\infty f(t)\,e^{-(s+a)t}\,dt = F(s+a). \quad ■ \end{aligned} \tag{3}$$

EXAMPLE 1. Determine $L\left\{t^3e^{5t}\right\}$. From Appendix C, $L\left\{t^3\right\} = 6/s^4$ so it follows from Theorem 5.7.1 that

$$L\left\{t^3e^{5t}\right\} = \frac{6}{(s-5)^4}. \quad ■ \tag{4}$$

EXAMPLE 2. We can invert $(2s+1)/(s^2+2s+4)$ by partial fractions, but it is simpler to note that

$$\begin{aligned} L^{-1}\left\{\frac{2s+1}{s^2+2s+4}\right\} &= L^{-1}\left\{\frac{2s+1}{(s+1)^2+3}\right\} = L^{-1}\left\{\frac{2(s+1)-1}{(s+1)^2+3}\right\} \\ &= 2L^{-1}\left\{\frac{(s+1)}{(s+1)^2+3}\right\} - L^{-1}\left\{\frac{1}{(s+1)^2+3}\right\} \\ &= 2e^{-t}\cos\sqrt{3}t - e^{-t}\frac{\sin\sqrt{3}t}{\sqrt{3}}, \end{aligned} \tag{5}$$

where in the last step we use entries 3 and 4 in Appendix C and Theorem 5.7.1. ■

THEOREM 5.7.2 *t-Shift*
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then for any constant $a > 0$

$$L\{H(t-a)f(t-a)\} = e^{-as}F(s) \tag{6}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\{e^{-as}F(s)\} = H(t-a)f(t-a). \tag{7}$$

Equations (6) and (7) are already given in Section 5.5, where we studied the Heaviside step function, but we repeat them here because the t-shift results seem a natural companion for the s-shift results given in Theorem 5.7.1.

THEOREM 5.7.3 *Multiplication by* $1/s$
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s} \tag{8}$$

for $s > \max\{0, s_0\}$ or, equivalently,

$$L^{-1}\left\{\frac{F(s)}{s}\right\} = \int_0^t f(\tau)\,d\tau. \tag{9}$$

Proof: This theorem is but a special case of the convolution theorem. Specifically, $\int_0^t f(\tau)\,d\tau = 1 * f$ so, according to that theorem,

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = L\{1 * f\} = L\{1\}\,L\{f\} = \frac{1}{s}F(s), \tag{10}$$

as asserted. ■

EXAMPLE 3. To evaluate $L^{-1}\{1/[s(s^2+1)]\}$, for example, we identify $F(s)$ as $1/(s^2+1)$. Since $f(t) = L^{-1}\{1/(s^2+1)\} = \sin t$,

$$L^{-1}\left\{\frac{1}{s(s^2+1)}\right\} = \int_0^t \sin\tau\,d\tau = 1 - \cos t. \tag{11}$$

Alternatively, we could have used partial fractions. ▮

Next, we obtain two useful theorems by differentiating and integrating the definition

$$F(s) = \int_0^\infty f(t)\,e^{-st}\,dt \tag{12}$$

with respect to s. First, we state without proof that if the integral in (12) converges for $s > s_0$, then

$$\begin{aligned}\frac{dF(s)}{ds} &= \frac{d}{ds}\int_0^{\infty} f(t)\,e^{-st}\,dt = \int_0^{\infty} \frac{\partial}{\partial s}\left[f(t)\,e^{-st}\right]\,dt \\ &= -\int_0^{\infty} t\,f(t)\,e^{-st}\,dt = -L\left\{t\,f(t)\right\},\end{aligned} \tag{13}$$

for $s > s_0$, and

$$\int_a^b F(s)\,ds = \int_a^b \int_0^{\infty} f(t)\,e^{-st}\,dt\,ds = \int_0^{\infty} f(t)\left(\int_a^b e^{-st}\,ds\right)\,dt \tag{14}$$

for $b \geq a \geq s_0$. The key step in (13) is the second equality, where we have inverted the order of the integration with respect to t and the differentiation with respect to s. In (14), the key is again the second equality, where we have inverted the order of integration with respect to t and the integration with respect to s.

In particular, if $a = s$ and $b = \infty$, then (14) becomes

$$\begin{aligned}\int_s^{\infty} F(s')\,ds' &= \int_0^{\infty} f(t)\left(\int_s^{\infty} e^{-s't}\,ds'\right)\,dt \\ &= \int_0^{\infty} \frac{f(t)}{t}\,e^{-st}\,dt = L\left\{\frac{f(t)}{t}\right\}\end{aligned} \tag{15}$$

for all $s > s_0$.

For reference, we state the results (13) and (15) as theorems.

THEOREM 5.7.4 *Differentiation with Respect to s*
If $L\{f(t)\} = F(s)$ exists for $s > s_0$, then

$$L\{t\,f(t)\} = -\frac{dF(s)}{ds} \tag{16}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\left\{\frac{dF(s)}{ds}\right\} = -t\,f(t). \tag{17}$$

EXAMPLE 4. From the known transform

$$L\{\sin at\} = \frac{a}{s^2 + a^2}, \qquad (s > 0) \tag{18}$$

we may use (16) to infer the additional results

$$L\{t\,\sin at\} = -\frac{d}{ds}\frac{a}{s^2 + a^2} = \frac{2as}{(s^2 + a^2)^2}, \qquad (s > 0) \tag{19}$$

$$L\left\{t^2 \sin at\right\} = -\frac{d}{ds}\frac{2as}{(s^2+a^2)^2} = 2a\,\frac{3s^2-a^2}{(s^2+a^2)^3}, \qquad (s>0) \tag{20}$$

and so on. ▮

THEOREM 5.7.5 *Integration with Respect to s*
If there is a real number s_0 such that $L\{f(t)\} = F(s)$ exists for $s > s_0$, and $\lim_{t\to 0} f(t)/t$ exists, then

$$L\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(s')\,ds' \tag{21}$$

for $s > s_0$ or, equivalently,

$$L^{-1}\left\{\int_s^\infty F(s')\,ds'\right\} = \frac{f(t)}{t}. \tag{22}$$

EXAMPLE 5. To evaluate

$$L^{-1}\left\{\ln\frac{s-a}{s-b}\right\}, \tag{23}$$

where a and b are real numbers, note that

$$\frac{d}{ds}\ln\frac{s-a}{s-b} = \frac{1}{s-a} - \frac{1}{s-b}$$

so

$$\ln\frac{s-a}{s-b} - \ln\frac{s_1-a}{s_1-b} = \int_{s_1}^{s}\left(\frac{1}{s'-a} - \frac{1}{s'-b}\right)ds' \tag{24}$$

for any s_1. Letting $s_1 \to \infty$ and recalling that $\ln 1 = 0$, (24) gives

$$\ln\frac{s-a}{s-b} = -\int_s^\infty\left(\frac{1}{s'-a} - \frac{1}{s'-b}\right)ds'. \tag{25}$$

Thus, identify $F(s)$ in (21) as $-1/(s-a) + 1/(s-b)$ (which does exist for $s > \max\{a,b\} \equiv s_0$). Then

$$f(t) = L^{-1}\left\{-\frac{1}{s-a} + \frac{1}{s-b}\right\} = e^{bt} - e^{at}. \tag{26}$$

Furthermore,

$$\lim_{t\to 0}\frac{f(t)}{t} = \lim_{t\to 0}\frac{e^{bt}-e^{at}}{t} = b - a \tag{27}$$

does exist, the last equality following from l'Hôpital's rule, so (22) gives the desired inverse as

$$L^{-1}\left\{\ln\frac{s-a}{s-b}\right\} = \frac{e^{bt}-e^{at}}{t}. \quad ▮ \tag{28}$$

THEOREM 5.7.6 *Large s Behavior of $F(s)$*
Let $f(t)$ be piecewise continuous on $0 \leq t \leq t_0$ for each finite t_0 and of exponential order as $t \to \infty$. Then
(i) $F(s) \to 0$ as $s \to \infty$,
(ii) $sF(s)$ is bounded as $s \to \infty$.

Proof: Since $f(t)$ is of exponential order as $t \to \infty$, there exist real constants K and c, with $K \geq 0$ such that $|f(t)| \leq K\,e^{ct}$ for all $t \geq t_0$ for some sufficiently large t_0. And since $f(t)$ is piecewise continuous on $0 \leq t \leq t_0$, there must be a finite constant M such that $|f(t)| \leq M$ on $0 \leq t \leq t_0$. Then

$$\begin{aligned}|F(s)| &= \left|\int_0^\infty f(t)\,e^{-st}\,dt\right| \leq \int_0^\infty |f(t)|\,e^{-st}\,dt \\ &= \int_0^{t_0} |f(t)|\,e^{-st}\,dt + \int_{t_0}^\infty |f(t)|\,e^{-st}\,dt \\ &\leq \int_0^{t_0} M\,e^{-st}\,dt + \int_{t_0}^\infty K\,e^{-(s-c)t}\,dt \\ &= M\frac{1-e^{-st_0}}{s} + K\frac{e^{-(s-c)t}}{-(s-c)}\bigg|_{t_0}^{\infty} < \frac{M}{s} + \frac{K}{s-c} \end{aligned} \tag{29}$$

for all $s > c$. It follows from this result that $F(s) \to 0$ as $s \to \infty$, and that $sF(s)$ is bounded as $s \to \infty$. ■

For instance, for each of the entries 1–7 in Appendix C we do have $F(s) \to 0$ and $sF(s)$ bounded as $s \to \infty$. For entry 8 we do too, unless $-1 < p < 0$, in which case $F(s) \to 0$ but $sF(s)$ is not bounded. However, in this case $f(t) = t^p$ is not piecewise continuous since $f(t) \to \infty$ as $t \to 0$ if p is negative.

THEOREM 5.7.7 *Initial-Value Theorem*
Let f be continuous and f' be piecewise continuous on $0 \leq t \leq t_0$ for each finite t_0, and let f and f' be of exponential order as $t \to \infty$. Then

$$\lim_{s\to\infty} [s\,F(s)] = f(0). \tag{30}$$

Proof: With the stated assumptions on f and f', it follows from Theorem 5.3.3 that

$$L\{f'(t)\} = s\,L\{f(t)\} - f(0). \tag{31}$$

Since f' satisfies the conditions of Theorem 5.5.7, it follows that $L\{f'(t)\} \to 0$ as $s \to \infty$. Thus, letting $s \to \infty$ in (31) gives the result stated in (30). ■

Normally, we invert $F(s)$ and obtain $f(t)$. However, if it is only $f(0)$ that we desire, not $f(t)$, we do not need to invert $F(s)$; all we need to do, according to (30), is to determine the limit of $sF(s)$ as $s \to \infty$.

As our final item, we show how to transform a periodic function, which is important in applications. First, we define what is meant by a periodic function. If there exists a positive constant T such that

$$f(t+T) = f(t) \tag{32}$$

for all $t \geq 0$, then we say that f is **periodic**, with **period** T.

EXAMPLE 6. The function $\sin t$ is periodic with period 2π since $\sin(t+2\pi) = \sin t \cos 2\pi + \sin 2\pi \cos t = \sin t$, for all t. ▮

EXAMPLE 7. The function f shown in Fig. 1 is, by inspection, seen to be periodic with period $T = 4$, for if we "stamp out" the segment ABC repeatedly, then we generate the graph of f. ▮

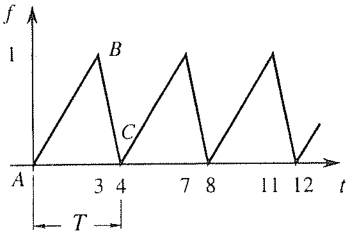


Figure 1. Periodic function f.

Notice that if f is periodic with period T, then it is also periodic with period $2T$, $3T$, $4T$, and so on. For instance, it follows from (32) that

$$f(t+2T) = f((t+T)+T) = f(t+T) = f(t)$$

so that if f is periodic with period T then it is also periodic with period $2T$. If there is a smallest period, it is called the **fundamental period**. Thus, $\sin t$ in Example 6 is periodic with period $2\pi, 4\pi, 6\pi, \ldots$, so its fundamental period is 2π; $f(t)$ in Example 7 is periodic with period $4, 8, 12, \ldots$, so its fundamental period is 4. In contrast, $f(t) = 3$ (i.e., a constant) is periodic with period T for *every* $T > 0$. Thus, there is no smallest period, and hence this f does not have a fundamental period.

To evaluate the Laplace transform of a periodic function $f(t)$, with period T (which is normally taken to be the fundamental period of f, if f has a fundamental period), it seems like a good start to break up the integral on t as

$$L\{f(t)\} = \int_0^\infty f(t)\, e^{-st}\, dt = \int_0^T f(t)\, e^{-st}\, dt + \int_T^{2T} f(t)\, e^{-st}\, dt + \cdots. \tag{33}$$

Next, let $\tau = t$ in the first integral on the right side of (33), $\tau = t - T$ in the second, $\tau = t - 2T$ in the third, and so on. Thus,

$$\begin{aligned} L\{f(t)\} = &\int_0^T f(\tau)\, e^{-s\tau}\, d\tau + \int_0^T f(\tau+T)\, e^{-s(\tau+T)}\, d\tau \\ &+ \int_0^T f(\tau+2T)\, e^{-s(\tau+2T)}\, d\tau + \cdots, \end{aligned} \tag{34}$$

but $f(\tau+T)=f(\tau)$, $f(\tau+2T)=f(\tau)$, and so on, so (34) becomes

$$L\{f(t)\}=\left(1+e^{-sT}+e^{-2sT}+\cdots\right)\int_0^T f(\tau)\,e^{-s\tau}\,d\tau. \tag{35}$$

Unfortunately, this expression is an infinite series. However, observe that

$$1+e^{-sT}+e^{-2sT}+\cdots=1+\left(e^{-sT}\right)+\left(e^{-2sT}\right)^2+\cdots \tag{36}$$

is a geometric series $1+z+z^2+\cdots$, with $z=e^{-sT}$, and the latter is known to have the sum $1/(1-z)$ if $|z|<1$. Since $|z|=\left|e^{-sT}\right|=e^{-sT}<1$ if $s>0$, we can sum the parenthetic series in (35) as $1/(1-e^{-sT})$.

Finally, if we ask that f be piecewise continuous on $0\le t\le T$, to ensure the existence of the integral in (35), then we can state the result as follows.

THEOREM 5.7.8 *Transform of Periodic Function*
If f is periodic with period T on $0\le t<\infty$ and piecewise continuous on one period, then

$$L\{f(t)\}=\frac{1}{1-e^{-sT}}\int_0^T f(t)\,e^{-st}\,dt \tag{37}$$

for $s>0$.

The point, of course, is that (37) requires integration only over one period, rather than over $0\le t<\infty$, and gives the transform in closed form rather than as an infinite series.

EXAMPLE 8. If f is the sawtooth wave shown in Fig. 2, then $T=2$, and

$$\int_0^T f(t)\,e^{-st}\,dt=\int_0^2 2t\,e^{-st}\,dt=2\frac{1-(1+2s)e^{-2s}}{s^2}, \tag{38}$$

so

$$L\{f(t)\}=\frac{1}{1-e^{-2s}}\,\frac{2\left[1-(1+2s)e^{-2s}\right]}{s^2}=\frac{2}{s^2}-\frac{4}{s}\,\frac{e^{-2s}}{1-e^{-2s}} \tag{39}$$

for $s>0$.

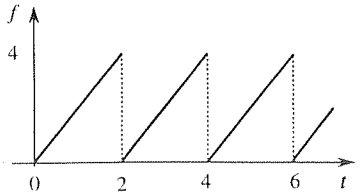


Figure 2. Sawtooth wave.

A more interesting question is the reverse: What is the inverse of

$$F(s)=\frac{2}{s^2}-\frac{4}{s}\,\frac{e^{-2s}}{1-e^{-2s}}, \tag{40}$$

where we pretend no advance knowledge of the sawtooth wave in Fig. 2, or even the knowledge that $f(t)$ is periodic? The key is to proceed in reverse – that is, to expand the $1/(1-e^{-2s})$ in a geometric series in powers of e^{-2s}. Thus

$$\begin{aligned}F(s)&=\frac{2}{s^2}-\frac{4}{s}e^{-2s}\left(1+e^{-2s}+e^{-4s}+\cdots\right)\\&=\frac{2}{s^2}-\frac{4}{s}\left(e^{-2s}+e^{-4s}+e^{-6s}+\cdots\right).\end{aligned} \tag{41}$$

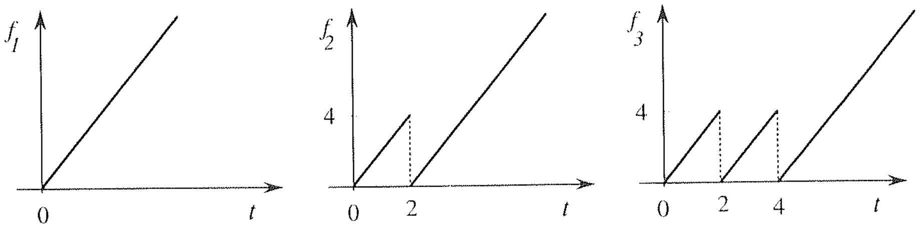


Figure 3. Partial sums of (42).

Assuming that the series can be inverted termwise,

$$f(t) = 2t - 4\left[H(t-2) + H(t-4) + H(t-6) + \cdots\right]. \tag{42}$$

The first few partial sums,

$$\begin{aligned} f_1(t) &= 2t, \\ f_2(t) &= 2t - 4H(t-2), \\ f_2(t) &= 2t - 4H(t-2) - 4H(t-4) \end{aligned}$$

are sketched in Fig. 3, and it is easy to infer that (42) gives the periodic sawtooth wave shown in Fig. 2.

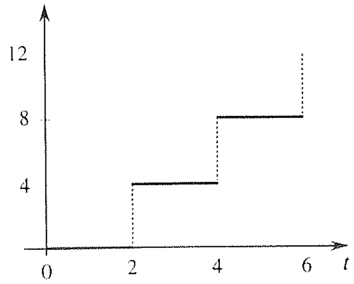


Figure 4. The staircase $4[H(t-2) + H(t-4) + \cdots]$.

COMMENT. Observe that the presence of $1 - e^{-sT}$ in the denominator of a transform does not suffice to imply that the inverse is periodic. For example, the inverse of $4e^{-2s}/[s(1 - e^{-2s})]$, in (40), is the nonperiodic "staircase" shown in Fig. 4. It is only when this staircase is subtracted from $2t$ that a periodic function results. ▮

This completes our discussion of the Laplace transform. Just as we used it, in this chapter, to solve linear ordinary differential equation initial-value problems, in later chapters we will use it to solve linear partial differential equation initial-value problems.

Closure. It would be difficult to pick out the one or two most important results in this section, since there are eight theorems, all of comparable importance. Most of these theorems are included as entries within Appendix C.

EXERCISES 5.7

1. Invert each of the following by any method. Cite any items from Appendix C and any theorems that you use. If it is applicable, verify the initial-value theorem, namely, that $sF(s) \to f(0)$ as $s \to \infty$; if it is not, then state why it is not.

(a) $\dfrac{1}{(s^2 + a^2)^2}$ (b) $\dfrac{1}{(s^2 - a^2)^2}$ (c) $\dfrac{s}{(s-2)^2}$

(d) $\dfrac{s^2}{(s+1)^3}$ (e) $\dfrac{1}{\sqrt{s+1}}$ (f) $\dfrac{s}{(s-a)^{3/2}}$

(g) $\dfrac{e^{-s}}{s^5}$ (h) $\dfrac{s}{(s+4)^6}$ (i) $\dfrac{e^{-2s}}{(s+1)^2}$

(j) $\ln\left(1+\dfrac{a^2}{s^2}\right)$ (k) $\dfrac{e^{-\pi s}}{(s+2)^4}$ (l) $\ln\left(1-\dfrac{a^2}{s^2}\right)$

(m) $\dfrac{e^{-s}}{s^2+s+1}$ (n) $\dfrac{e^{-3s}}{s^2+2s-4}$ (o) $\dfrac{e^{-2s}}{s^2(s^2-2s-2)}$

(p) $\ln\left(\dfrac{s^2+1}{s^2+s}\right)$ (q) $\dfrac{1}{s(1-e^{-s})}$ (r) $\dfrac{1}{s(1+e^{-s})}$

(s) $\dfrac{1}{s}\tanh s$ (t) $\dfrac{1}{s^2(1-e^{-s})}$ (u) $\dfrac{1}{(s+1)(1-e^{-2s})}$

(v) $\dfrac{e^{-s}}{s-8}$ (w) $\dfrac{s}{(s^2+1)^{3/2}}$ (x) $\dfrac{1}{s(1-e^{s})}$

2 (a)–(x) Invert the transform given in the corresponding part of Exercise 1, using computer software.

3. (a) In the simple case where $f(t) = \sin t$, show that (37) does indeed give

$$L\{\sin t\} = \frac{1}{s^2+1}.$$

(b) In the case where $f(t) = \cos t$, show that (37) does indeed give

$$L\{\cos t\} = \frac{s}{s^2+1}.$$

4. (*Scale changes*) Show that if $L\{f(t)\} = F(s)$, then

(a) $L\{f(at)\} = \dfrac{1}{a}F\left(\dfrac{s}{a}\right)$ (b) $L^{-1}\{F(as)\} = \dfrac{1}{a}f\left(\dfrac{t}{a}\right)$

5. Determine the Laplace transform of the function $f(t)$ that is periodic and defined on one period as follows.

(a) $\sin t, \quad 0 \le t \le \pi$ (b) $\begin{cases} 1, & 0 \le t < 2 \\ 0, & 2 \le t < 3 \end{cases}$

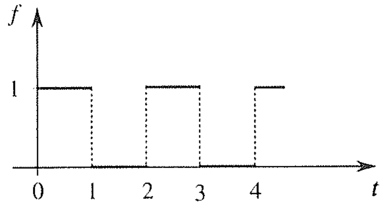
(c) $\sin 2t, \quad 0 \le t \le \pi$ (d) $e^{-t}, \quad 0 \le t < 2$

(e) $\begin{cases} t, & 0 \le t < 1 \\ 0, & 1 \le t < 2 \end{cases}$ (f) $\begin{cases} t, & 0 \le t < 1 \\ t-2, & 1 \le t < 2 \end{cases}$

(g) $\begin{cases} 2, & 0 \le t < 1 \\ 4, & 1 \le t < 2 \\ 1, & 2 \le t < 3 \end{cases}$

6. (a) Solve $x' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$ and $f(t)$ is the square wave shown, and show that the solution is

$$\begin{aligned} x(t) &= x_0 e^{-t} + \left[1 - e^{-(t-0)}\right] H(t-0) \\ &\quad - \left[1 - e^{-(t-1)}\right] H(t-1) \\ &\quad + \left[1 - e^{-(t-2)}\right] H(t-2) - \cdots . \end{aligned} \tag{6.1}$$

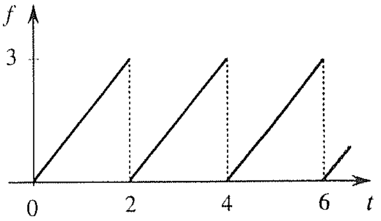


HINT: It would be wasteful to determine $F(s)$ because the solution can be expressed as a convolution integral involving $f(t)$ directly. In that integral, express $f(t)$ as an infinite series of Heaviside functions.

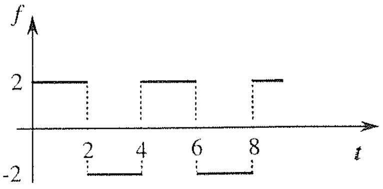
(b) Sketch and label the graph of $x(t)$ over $0 < t < 3$, say, for the case where $x_0 = 0$. Is $x(t)$ periodic? If not, is there a value of x_0 such that $x(t)$ is periodic? Explain.

7. Solve $x' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$ and $f(t)$ is the periodic function shown. HINT: Read Exercise 6.

(a)



(b)



8. Solve $x'' + x = f(t)$ by the Laplace transform, where $x(0) = x_0$, $x'(0) = x_0'$, and $f(t)$ is the square wave shown in Exercise 6. Evaluate $x(5)$ if $x_0 = x_0' = 1$.

Chapter 5 Review

The Laplace transform has a variety of uses, but its chief application is in the solution of linear ordinary and partial differential differential equations. In this chapter our focus is on its use in solving linear ordinary differential equations with constant coefficients. The power of the method is that it reduces such a differential equation, homogeneous or not, to a linear algebraic one. The hardest part, the inversion, is often accomplished with the help of tables, a considerable number of theorems, and computer software. Also, any initial conditions that are given become built in, in the process of transforming the differential equation, so they do not need to be applied at the end.

Chief properties given in Section 5.3 are:

Linearity of the transform and its inverse

$$L\left\{\alpha u(t)+\beta v(t)\right\}=\alpha\, L\left\{u(t)\right\}+\beta\, L\left\{v(t)\right\},$$
$$L^{-1}\left\{\alpha U(s)+\beta V(s)\right\}=\alpha\, L^{-1}\left\{U(s)\right\}+\beta\, L^{-1}\left\{V(s)\right\},$$

Transform of derivatives

$$L\left\{f'\right\}=sF(s)-f(0),\quad L\left\{f''\right\}=s^2F(s)-sf(0)-f'(0),\quad \cdots,$$

Convolution Theorem

$$L\left\{(f*g)(t)\right\}=F(s)G(s)$$

or

$$L^{-1}\left\{F(s)G(s)\right\}=(f*g)(t),$$

where

$$(f*g)(t)=\int_0^t f(\tau)\, g(t-\tau)\, d\tau$$

is the Laplace convolution of f and g.

In Sections 5.5 and 5.6 we introduce the step and impulse functions $H(t-a)$ and $\delta(t-a)$, defined by

$$H(t-a)=\begin{cases}0, & t<a\\ 1, & t\geq a\end{cases}$$

and

$$\int_A^B g(t)\,\delta(t-a)\,dt=\begin{cases}g(a), & A\leq a<B\\ 0, & a<A \quad\text{or}\quad a\geq B,\end{cases}$$

to model piecewise-defined and impulsive forcing functions.

Finally, in Section 5.7 we derive additional properties:

s-shift

$$L\left\{e^{-at}f(t)\right\}=F(s+a)$$

or

$$L^{-1}\left\{F(s+a)\right\} = e^{-at}f(t).$$

t-shift

$$L\left\{H(t-a)f(t-a)\right\} = e^{-as}F(s)$$

or

$$L^{-1}\left\{e^{-as}F(s)\right\} = H(t-a)f(t-a).$$

Multiplication by $1/s$

$$L\left\{\int_0^t f(\tau)\,d\tau\right\} = \frac{F(s)}{s}$$

or

$$L^{-1}\left\{\frac{F(s)}{s}\right\} = \int_0^t f(\tau)\,d\tau.$$

Differentiation with respect to s

$$L\left\{t\,f(t)\right\} = -\frac{dF(s)}{ds}$$

or

$$L^{-1}\left\{\frac{dF(s)}{ds}\right\} = -t\,f(t).$$

Integration with respect to s

$$L\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(s')ds'$$

or

$$L^{-1}\left\{\int_s^\infty F(s')ds'\right\} = \frac{f(t)}{t}.$$

Large s behavior of $F(s)$

$$F(s) \to 0 \quad \text{as} \quad s \to \infty,$$
$$sF(s) \quad \text{bounded as} \quad s \to \infty.$$

Transform of periodic function of period T

$$L\left\{f(t)\right\} = \frac{1}{1-e^{-sT}}\int_0^T f(t)e^{-st}dt.$$

NOTE: The preceding list is intended as an overview so, for brevity, the various conditions under which these results hold have been omitted.

Chapter 6

Quantitative Methods: Numerical Solution of Differential Equations

6.1 Introduction

Following the introduction in Chapter 1, Chapters 2–5 cover both the underlying theory of differential equations and analytical solution techniques as well. That is, the objective thus far has been to find an analytical solution – in closed form if possible or as an infinite series if necessary. Unfortunately, a great many differential equations encountered in applications, and most nonlinear equations in particular, are simply too difficult for us to find analytical solutions.

Thus, in Chapters 6 and 7 our approach is fundamentally different, and complements the analytical approach adopted in Chapters 2–5: in Chapter 6 we develop *quantitative* methods, and in Chapter 7 our view is essentially *qualitative*. More specifically, in this chapter we "discretize" the problem and seek, instead of an analytical solution, the numerical values of the dependent variable at a discrete set of values of the independent variable so that the result is a table or graph, with those values determined approximately (but accurately), rather than exactly.

Perhaps the greatest drawback to numerical simulation is that whereas an analytical solution explicitly displays the dependence of the dependent variable(s) on the various physical parameters (such as spring stiffnesses, driving frequencies, electrical resistances, inductances, and so on), one can carry out a numerical solution only for a specific set of values of the system parameters. Thus, parametric studies (i.e., studies of the qualitative and quantitative effects of the various parameters upon the solution) can be tedious and unwieldy, and it is useful to reduce the number of parameters as much as possible (by nondimensionalization, as discussed in Section 2.4.4) before embarking upon a numerical study.

The numerical solution of differential equations covers considerable territory so the present chapter is hardly complete. Rather, we aim at introducing the fundamental ideas, concepts, and potential difficulties, as well as specific methods that are accurate and readily implemented. We do mention computer software that carries out these computations automatically, but our present aim is to provide enough information so that you will be able to select a specific method and program it. In contrast, in Chapter 7, where we look more at qualitative issues, we rely heavily upon available software.

6.2 Euler's Method

In this section and the two that follow, we study the numerical solution of the first-order initial-value problem

$$y' = f(x, y); \quad y(a) = b \tag{1}$$

on $y(x)$.

To motivate the first and simplest of these methods, Euler's method, consider the problem

$$y' = y + 2x - x^2; \quad y(0) = 1 \quad (0 \leq x < \infty) \tag{2}$$

with the exact solution (Exercise 1)

$$y(x) = x^2 + e^x. \tag{3}$$

Of course, in practice one wouldn't solve (2) numerically because we can solve it analytically and obtain the solution (3), but we will use (2) as an illustration.

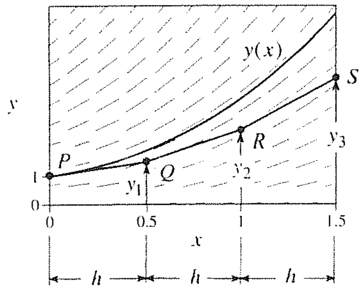


Figure 1. Direction field motivation of Euler's method, for the initial-value problem (2).

In Fig. 1 we display the direction field defined by $f(x, y) = y + 2x - x^2$, as well as the exact solution (3). In graphical terms, Euler's method amounts to using the direction field as a road map in developing an approximate solution to (2). Beginning at the initial point P, namely $(0, 1)$, we move in the direction dictated by the lineal element at that point. As seen from the figure, the farther we move along that line, the more we expect our path to deviate from the exact solution. Thus, the idea is not to move very far. Stopping at $x = 0.5$, say, for the sake of illustration, we revise our direction according to the slope of the lineal element at that point Q. Moving in that new direction until $x = 1$, we revise our direction at R, and so on, moving in x increments of 0.5. We call the x increment the **step size** and denote it as h. In Fig. 1, h is 0.5.

Let us denote the y values at $Q, R, \ldots$ as $y_1, y_2, \ldots$. They are computed as $y_1 = y_0 + f(x_0, y_0)h$, $y_2 = y_1 + f(x_1, y_1)h, \ldots$, where (x_0, y_0) is the initial point P. Expressed as a numerical algorithm, the **Euler method** is therefore as follows:

$$\boxed{y_{n+1} = y_n + f(x_n, y_n)h, \qquad (n = 0, 1, 2, \ldots)} \tag{4}$$

where f is the function on the right side of the given differential equation (1), $x_0 = a$, $y_0 = b$, h is the chosen step size, and $x_n = x_0 + nh$.

Euler's method is also known as the **tangent-line method** because the first straight-line segment of the approximate solution is tangent to the exact solution

$y(x)$ at P, and each subsequent segment emanating from (x_n, y_n) is tangent to the solution curve through that point.

Apparently, the greater the step size the less accurate the results, in general. For instance, the first point Q deviates more and more from the exact solution as the step size is increased – that is, as the segment PQ is extended. Conversely, we expect the approximate solution to approach the exact solution curve as h is reduced. This expectation is supported by the results shown in Table 1 for the initial-value problem (2), obtained by Euler's method with step sizes of $h = 0.5, 0.1$, and 0.02; we have included the exact solution $y(x)$, given by (3), for comparison. With $h = 0.5$, for instance,

Table 1. Comparison of numerical solution of (2) using Euler's method, with the exact solution.

x	$h = 0.5$	$h = 0.1$	$h = 0.02$	$y(x)$
0.5	1.5000	1.7995	1.8778	1.8987
1.0	2.6250	3.4344	3.6578	3.7183
1.5	4.4375	6.1095	6.5975	6.7317

$$\begin{aligned} y_1 &= y_0 + \left(y_0 + 2x_0 - x_0^2\right) h = 1 + (1 + 0 - 0)(0.5) = 1.5, \\ y_2 &= y_1 + \left(y_1 + 2x_1 - x_1^2\right) h = 1.5 + (1.5 + 1 - 0.25)(0.5) = 2.625, \\ y_3 &= y_2 + \left(y_2 + 2x_2 - x_2^2\right) h = 2.625 + (2.625 + 2 - 1)(0.5) = 4.4375. \end{aligned}$$

With $h = 0.1$, the values tabulated at $x = 0.5, 1.0, 1.5$ are y_5, y_{10}, y_{15}, with the intermediate computed y values omitted for brevity.

Scanning each row of the tabulation, we can see that the approximate solution appears to be converging to the exact solution as $h \to 0$ (though we cannot be certain from such results no matter how small we make h), and that the convergence is not very rapid, for even with $h = 0.02$ the computed value at $x = 1.5$ is in error by 2%.

As strictly computational as this sounds, two important theoretical questions present themselves: Does the method converge to the exact solution as $h \to 0$ and, if so, how fast? By a method being **convergent** we mean that for any fixed x value in the x interval of interest the sequence of y values, obtained using smaller and smaller step size h, tends to the exact solution $y(x)$ as $h \to 0$.

Let us see whether the Euler method is convergent. Observe that there are two sources of error in the numerical solution. One is the tangent-line approximation upon which the method is based, and the other is the accumulation of numerical roundoff errors within the computing machine since a machine can carry only a finite number of significant figures, after which it rounds off (or chops off, depending upon the machine). In discussing convergence, one ignores the presence of such roundoff error and considers it separately. Thus, in this discussion we imagine our computer to be perfect, carrying an infinite number of significant figures.

Local truncation error. Although we are interested in the accumulation of error after a great many steps have been carried out, to reach any given x, it seems best to begin by investigating the error incurred in a single step, from x_{n-1} to x_n (or from x_n to x_{n+1}, it doesn't matter). We need to distinguish between the exact and approximate solutions so let us denote the exact solution at x_n as $y(x_n)$ and the approximate numerical solution at x_n as y_n. These are given by the Taylor series

$$\begin{aligned} y(x_n) &= y(x_{n-1}) + y'(x_{n-1})\,(x_n - x_{n-1}) + \frac{y''(x_{n-1})}{2!}\,(x_n - x_{n-1})^2 + \cdots \\ &= y(x_{n-1}) + y'(x_{n-1})h + \frac{y''(x_{n-1})}{2!}h^2 + \cdots \end{aligned} \tag{5}$$

and the Euler algorithm

$$y_n = y_{n-1} + f\,(x_{n-1}, y_{n-1})\,h, \tag{6}$$

respectively. It is important to understand that the Euler method (6) amounts to retaining only the first two terms of the Taylor series in (5). Thus, it replaces the actual function by its tangent-line approximation.

We suppose that $y(x_{n-1})$ and y_{n-1} are identical, and we ask how large the error $e_n \equiv y(x_n) - y_n$ is after making that single step, from x_{n-1} to x_n. We can get an expression for e_n by subtracting (6) from (5), but the right side will be an infinite series. Thus, it is more convenient to use, in place of the (infinite) Taylor series (5), the (finite) Taylor's formula with remainder,

$$y(x_n) = y(x_{n-1}) + y'(x_{n-1})h + \frac{y''(\xi)}{2!}h^2, \tag{7}$$

where ξ is some point in the interval $[x_{n-1}, x_n]$. Now, subtracting (6) from (7), and noting that $y'(x_{n-1}) = f\,[x_{n-1}, y(x_{n-1})] = f(x_{n-1}, y_{n-1})$ because of our supposition that $y(x_{n-1}) = y_{n-1}$, gives

$$e_n = \frac{y''(\xi)}{2}h^2. \tag{8}$$

The latter expression for e_n is awkward to apply since we don't know ξ, except that $x_{n-1} \le \xi \le x_n$.* However, (8) is of more interest in that it shows how the single-step error e_n varies with h. Specifically, since $x_{n-1} \le \xi \le x_{n-1} + h$, we see that as $h \to 0$ we have $\xi \to x_{n-1}$, so (8) gives $e_n \sim \dfrac{y''(x_{n-1})}{2}h^2 = Ch^2$ as $h \to 0$, where C is a constant. Accordingly, we say that e_n is of order h^2 and write

$$\boxed{e_n = O(h^2)} \tag{9}$$

*It also appears that we do not know the y'' function, but it follows from (2) that $y'' = y' + 2 - 2x = (y + 2x - x^2) + 2 - 2x = y + 2 - x^2$.

as $h \to 0$. [The big oh notation is defined in Section 4.5, and (9) simply means that $e_n \sim Ch^2$ as $h \to 0$ for some nonzero constant C.]

Since the error e_n is due to truncation of the Taylor series it is called the truncation error–more specifically, the **local truncation error** because it is the truncation error incurred in a single step.

Accumulated truncation error and convergence. Of ultimate interest, however, is the truncation error that has accumulated over *all* of the preceding steps since that error is the difference between the exact solution and the computed solution at any given x_n. We denote it as $E_n \equiv y(x_n) - y_n$ and call it the **accumulated truncation error**. If it seems strange that we have defined both the local and accumulated truncation errors as $y(x_n) - y_n$, it must be remembered that the former is that which results from a single step (from x_{n-1} to x_n) whereas the latter is that which results from the entire sequence of steps (from x_0 to x_n).

We can estimate E_n at a fixed x location (at least insofar as its order of magnitude) as the local truncation error e_n times the number of steps n. Since $e_n = O(h^2)$, this idea gives

$$E_n = O(h^2) \cdot n = O(h^2)\frac{n\,h}{h} = O(h^2)\frac{x_n}{h} = O(h) \cdot x_n = O(h), \qquad (10)$$

where the last step follows from the fact that the selected location x_n is held fixed as $h \to 0$. Since the big oh notation is insensitive to scale factors, the x_n factor can be absorbed by the $O(h)$ so

$$\boxed{E_n = O(h),} \qquad (11)$$

which result tells us how fast the numerical solution converges to the exact solution (at any fixed x location) as $h \to 0$. Namely, $E_n \sim Ch$ for some nonzero constant C. To illustrate, consider the results shown in Table 1, and consider $x = 1.5$, say, in particular. According to $E_n \sim Ch$, if we reduce h by a factor of five, from 0.1 to 0.02, then likewise we should reduce the error by a factor of five. We find that $(6.7317 - 6.1095)/(6.7317 - 6.5875) \approx 4.6$, which is indeed close to five. We can't expect it to be exactly five for two reasons: First, (11) holds only as $h \to 0$, whereas we have used $h = 0.1$ and 0.02. Second, we obtained the values in Table 1 using a computer, and a computer introduces an additional error, due to roundoff, which has not been accounted for in our derivation of (11). Probably it is negligible in this example.

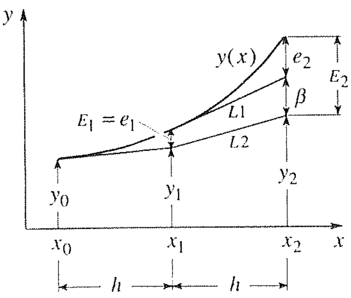


Figure 2. The global truncation error.

While (11) can indeed be proved rigorously, be aware that our reasoning in (10) was only heuristic. To understand the shortcoming of our logic, consider the diagram in Fig. 2, where we show only two steps, for simplicity.

Our reasoning, in writing $E_n = O(h^2) \cdot n$ in (10), is that the accumulated truncation error E_n is (at least insofar as order of magnitude) the sum of the n single-step errors. However, that is not quite true. We see from Fig. 2 that E_2 is $e_2 + \beta$, not the sum of the single-step errors $e_2 + e_1$, and β is not identical to e_1. The difference between β and e_1 is the result of the *slightly* different slopes of $L1$ and $L2$ acting over the *short* distance h, and that difference can be shown to be a higher-order effect that does not invalidate the final result that $E_n = O(h)$, provided that

f is well enough behaved (for example, if f, f_x, and f_y are all continuous on the x, y region of interest).

In summary, (11) shows that the Euler method (4) is convergent because the accumulated truncation error tends to zero as $h \to 0$. More generally if, for a given method, $E_n = O(h^p)$ as $h \to 0$, then the method is convergent if $p > 0$, and we say that it is **of order p**. Thus, *the Euler method is a first-order method.*

Although convergent and easy to implement, Euler's method is usually too inaccurate for serious computation because it is only a first-order method. That is, since the accumulated truncation error is proportional to h to the first power, we need to make h extremely small if the error is to be extremely small. Why can't we do that? Why can't we merely let $h = 10^{-8}$, say? There are two important reasons. One is that with $h = 10^{-8}$, it would take 10^8 steps to generate the Euler solution over a unit x interval. That number of steps might simply be impractically large in terms of computation time and expense.

Second, besides the truncation error that we have discussed there is also machine roundoff error, and that error can be expected to grow with the number of calculations. Thus, as we diminish the step size h and increase the number of steps, to reduce the truncation error, we inflict a roundoff error penalty that diminishes the intended increase in accuracy. In fact, we can anticipate the existence of an optimal h value so that to decrease h below that value is counterproductive. Said differently, a given level of accuracy may prove unobtainable because of the growth in the roundoff error as h is reduced. Further discussion of this point is contained in the exercises.

Finally, there is an important practical question not yet addressed: How do we know how small to choose h? We will have more to say about this later, but for now let us give a simple procedure, namely, reducing h until the results settle down to the desired accuracy. For instance, suppose we solve (2) by Euler's method using $h = 0.5$, as a first crack. Pick any fixed point x in the interval of interest, such as $x = 1.5$. The computed solution there is 4.4375. Now reduce h, say to 0.01, and run the program again. The result this time, at $x = 1.5$, is 6.1095. Since those results differ considerably, reduce h again, say to 0.02, and run the program again. Simpy repeat that procedure until the solution at $x = 1.5$ settles down to the desired number of significant figures. Accept the results of the final run, and discard the others. (Of course, one will not have an exact solution to compare with as we did in Table 1.)

The foregoing idea is merely a rule of thumb, and is the same idea that we use in computing an infinite series: keep adding more and more terms until successive partial sums agree with the desired number of significant figures.

Closure. The Euler method is embodied in (4). It is easy to implement, either using a hand-held calculator or programming it to be run on a computer. The method is convergent but only of first order and hence is not very accurate. Thus, it is important to develop more accurate methods, and we do that in the next section.

We also use our discussion of the Euler method to introduce the concept of the local and accumulated truncation errors e_n and E_n, respectively, which are due to the approximate discretization of the problem and which have nothing to do with additional errors that enter due to machine roundoff. The former is the error incurred in a single step, and the latter is the accumulated error over the entire caclulation. Finally, we define the method to be convergent if the accumulated truncation error E_n tends to zero at any given fixed point x, as the step size h tends to zero, and of order p if $E_n = O(h^p)$ as $h \to 0$. The Euler method is convergent and of order one.

EXERCISES 6.2

1. Derive the particular solution (3) of the initial-value problem (2).

2. Use the Euler method to compute, by hand, y_1, y_2, and y_3 for the specified initial-value problem using $h = 0.2$.

(a) $y' = -y;\quad y(0) = 1$
(b) $y' = 2xy;\quad y(0) = 0$
(c) $y' = 3x^2y^2;\quad y(0) = 0$
(d) $y' = 1 + 2xy^2;\quad y(1) = -2$
(e) $y' = 2xe^{-y};\quad y(1) = -1$
(f) $y' = x^2 - y^2;\quad y(3) = 5$
(g) $y' = x \sin y;\quad y(0) = 0$
(h) $y' = \tan(x + y);\quad y(1) = 2$
(i) $y' = 5x - 2\sqrt{y};\quad y(0) = 4$
(j) $y' = \sqrt{x + y};\quad y(0) = 3$

3. Program and run Euler's method for the initial-value problem $y' = f(x, y)$, with $y(0) = 1$ and $h = 0.1$, through y_{10}. Print $y_1, \dots, y_{10}$ and the exact solution $y(x_1), \dots, y(x_{10})$ as well. (Six significant figures will suffice.) Evaluate E_{10}. Use the $f(x, y)$ specified below.

(a) $2x$ (b) $-6y^2$ (c) $x + y$
(d) $y \sin x$ (e) $(y^2 + 1)/2$ (f) $4xe^{-y}$
(g) $1 + x^2 + y$ (h) $-y \tan x$ (i) e^{x+y}

4. (a)–(h) Program and run Euler's method for the initial-value problem $y' = f(x, y)$ (with f given in the corresponding part of Exercise 3), and print out the result at $x = 0.5$. Use $h = 0.1$, then 0.05, then 0.01, then 0.005, then 0.001, and compute the accumulated truncation error at $x = 0.5$ for each case. Is the rate of decrease of the accumulated truncation error, as h decreases, consistent with the fact that Euler's method is a first-order method? Explain.

5. Thus far we have taken the step h to be positive, and therefore developed a solution to the right of the initial point. Is Euler's method valid if we use a negative step, $h < 0$, and hence develop a solution to the left? Explain.

6. We have seen that by discretizing the problem, we can approximate the solution $y(x)$ of a differential equation $y' = f(x, y)$ by a discrete variable y_n by solving

$$y_{n+1} = y_n + f(x_n, y_n)h \tag{6.1}$$

sequentially, for $n = 0, 1, 2, \dots$. Besides being a numerical algorithm for the calculation of the y_n's, (6.1) is an example of a **difference equation** governing the sequence of y_n's, just as $y' = f(x, y)$ is a differential equation governing $y(x)$. If f is simple enough it may be possible to solve (6.1) for y_n analytically, and that idea is the focus of this exercise. Specifically, consider $y' = Ay$, where A is a given constant. Then (6.1) becomes

$$y_{n+1} = (1 + Ah)y_n. \tag{6.2}$$

(a) Derive the solution

$$y_n = C(1 + Ah)^n \tag{6.3}$$

of (6.2), where C is the initial value y_0, if one is specified.
(b) Show that as $h \to 0$ (6.3) does converge to the solution Ce^{Ax} of the original equation $y' = Ay$. HINT: Begin by expressing $(1 + Ah)^n$ as $e^{\ln(1+Ah)^n}$. NOTE: Thus, for the simple differential equation $y' = Ay$ we have been able to prove the convergence of the Euler method by actually solving (6.2) for y_n, in closed form, then taking the limit of that result as $h \to 0$.
(c) Use computer software to obtain the solution (6.3) of the difference equation (6.2). On *Maple*, for instance, use the rsolve command.

7. In this section we have taken the step size h to be a constant from one step to the next. Is there any reason why we could not vary h from one step to the next? Explain.

6.3 Improvements: Midpoint Rule and Runge–Kutta

Our objective in this section is to develop more accurate methods than the first-order Euler method – namely, higher-order methods. In particular, we are aiming at the widely used fourth-order Runge–Kutta method, which is an excellent general-purpose differential equation solver. To bridge the gap between these two methods, we begin with some general discussion about how to develop higher-order methods.

6.3.1. Midpoint rule. To derive more accurate differential equation solvers, Taylor series (better yet, Taylor's formula with remainder) offers a useful line of approach. To illustrate, consider the Taylor's formula with remainder,

$$y(x) = y(a) + y'(a)(x-a) + \frac{y''(\xi)}{2!}(x-a)^2, \tag{1}$$

where ξ is some point in $[a,x]$. If we let $x = x_{n+1}$, $a = x_n$, and $x - a = x_{n+1} - x_n = h$, then (1) becomes

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(\xi)}{2!}h^2. \tag{2}$$

Since $y' = f(x,y)$, we can replace the $y'(x_n)$ in (2) by $f(x_n, y(x_n))$. Also, the last term in (2) can be expressed more simply as $O(h^2)$ so we have

$$y(x_{n+1}) = y(x_n) + f(x_n, y(x_n))h + O(h^2). \tag{3}$$

If we neglect the $O(h^2)$ term and call attention to the approximation thereby incurred by replacing the exact values $y(x_{n+1})$ and $y(x_n)$ by the approximate values y_{n+1} and y_n, respectively, then we have the Euler method

$$y_{n+1} = y_n + f(x_n, y_n)h. \tag{4}$$

Since the term that we dropped in passing from (3) to (4) was $O(h^2)$, the local truncation error is $O(h^2)$, and the accumulated truncation error is $O(h)$.

One way to obtain a higher-order method is to retain more terms in the Taylor's formula. For instance, begin with

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \frac{y'''(\eta)}{6}h^3 \tag{5}$$

in place of (2) or, since $y'' = \dfrac{d}{dx}f(x,y(x)) = f_x + f_y y' = f_x + f_y f$,

$$\begin{aligned} y(x_{n+1}) &= y(x_n) + f(x_n, y(x_n))h \\ &\quad + \frac{1}{2}\left[f_x(x_n, y(x_n)) + f_y(x_n, y(x_n))f(x_n, y(x_n))\right]h^2 + O(h^3). \end{aligned} \tag{6}$$

If we truncate (6) by dropping the $O(h^3)$ term, and change $y(x_{n+1})$ and $y(x_n)$ to y_{n+1} and y_n, respectively, then we have the method

$$y_{n+1} = y_n + f(x_n, y_n)h + \frac{1}{2}\left[f_x(x_n, y_n) + f_y(x_n, y_n)f(x_n, y_n)\right] h^2 \qquad (7)$$

with a local truncation error that is $O(h^3)$ and an accumulated truncation error that is $O(h^2)$; that is, we now have a *second*-order method.

Why do we say that (7) is a second-order method? Following the same heuristic reasoning as in Section 6.2, the accumulated truncation error E_n is of the order of the local truncation error times the number of steps so

$$E_n = O(h^3) \cdot n = O(h^3)\frac{n\,h}{h} = O(h^3)\frac{x_n}{h} = O(h^2) \cdot x_n = O(h^2),$$

as claimed. In fact, as a simple rule of thumb it can be said that if the local truncation error is $O(h^p)$, with $p > 1$, then the accumulated truncation error is $O(h^{p-1})$, and one has a $(p-1)$th-order method.

Although the second-order convergence of (7) is an improvement over the first-order convergence of Euler's method, the attractiveness of (7) is diminished by an approximately threefold increase in the computing time per step since it requires three function evaluations (f, f_x, f_y) per step whereas Euler's method requires only one (f). It's true that to carry out one step of Euler's method we need to evaluate f, multiply that by h, and add the result to y_n, but we can neglect the multiplication by h and addition of y_n on the grounds that a typical $f(x, y)$ involves many more arithmetic steps than that. Thus, as a rule of thumb, one compares the computation time per step of two methods by comparing only the number of function evaluations per step.

Not satisfied with (7) because it requires three function evaluations, let us return to Taylor's formula (5). If we replace h by $-h$, that amounts to making a backward step so the term on the left will be $y(x_{n-1})$ instead of $y(x_{n+1})$. Making those changes, and also copying (5), for comparison, we have

$$y(x_{n-1}) = y(x_n) - y'(x_n)h + \frac{y''(x_n)}{2}h^2 - \frac{y'''(\zeta)}{6}h^3, \qquad (8a)$$

$$y(x_{n+1}) = y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \frac{y'''(\eta)}{6}h^3, \qquad (8b)$$

respectively, where ζ is some point in $[x_{n-1}, x_n]$ and η is some point in $[x_n, x_{n+1}]$. Now we can eliminate the bothersome y'' terms by subtracting (8a) from (8b). Doing so gives

$$y(x_{n+1}) - y(x_{n-1}) = 2y'(x_n)h + \frac{y'''(\eta) + y'''(\zeta)}{6}h^3$$

or

$$y(x_{n+1}) = y(x_{n-1}) + 2f\left(x_n, y(x_n)\right)h + O\left(h^3\right).$$

Finally, if we drop the $O\left(h^3\right)$ term and change $y(x_{n+1})$, $y(x_{n-1})$, $y(x_n)$ to y_{n+1}, y_{n-1}, y_n, respectively, we have

$$\boxed{y_{n+1} = y_{n-1} + f(x_n, y_n)(2h),} \tag{9}$$

which method is known as the **midpoint rule**. Like (7), the midpoint rule is a second-order method but, unlike (7), it requires only one function evaluation per step. It is an example of a **multi-step method** because it uses information from more than one of the preceding points – namely, from two: the nth and $(n-1)$th. Thus, it is a two-step method whereas Euler's method and the Taylor series method given by (7) are single-step methods.

A disadvantage of the midpoint rule (and other multi-step methods) is that it is not self-starting. That is, the first step gives y_1 in terms of x_0, y_0, y_{-1}, but y_{-1} is not defined. Thus, (9) applies only for $n \geq 1$, and to get the method started we need to compute y_1 by a different method. For instance, we could use Euler's method to obtain y_1 and then switch over to the midpoint rule (9). Of course, if we do that we should do it not in a single Euler step but in many so as not to degrade the subsequent second-order accuracy.

EXAMPLE 1. Consider the same "test problem" as in Section 6.2,

$$y' = y + 2x - x^2; \qquad y(0) = 1, \qquad (0 \leq x < \infty) \tag{10}$$

with the exact solution $y(x) = x^2 + e^x$. Let us use the midpoint rule with $h = 0.1$. To get it started, carry out ten steps of Euler's method with $h = 0.01$. The result of those steps is the approximate solution 1.11358 at $x = 0.1$, which we now take as y_1. Then proceeding with the midpoint rule we obtain from (9)

$$\begin{aligned} y_2 &= y_0 + 2\left(y_1 + 2x_1 - x_1^2\right) h \\ &= 1 + 2\,(1.11358 + 0.2 - 0.01)\,(0.1) = 1.26072 \\ y_3 &= y_1 + 2\left(y_2 + 2x_2 - x_2^2\right) h \\ &= 1.11358 + 2\,(1.26072 + 0.4 - 0.04)\,(0.1) = 1.43772, \end{aligned}$$

and so on. The results are shown in Table 1 and contrasted with the less accurate Euler results using the same step size, $h = 0.1$. ∎

Before leaving the midpoint rule, it is interesting to interpret the improvement in accuracy, from Euler to midpoint, graphically. If we solve

$$y(x_{n+1}) \approx y(x_n) + y'(x_n)h \qquad \text{(Euler)} \tag{11}$$

and

$$y(x_{n+1}) \approx y(x_{n-1}) + 2y'(x_n)h \qquad \text{(midpoint)} \tag{12}$$

for $y'(x_n)$, we have

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_n)}{h} \qquad \text{(Euler)} \tag{13}$$

Table 1. Comparison of Euler, midpoint rule, and exact solutions of the initial-value problem (10), with $h = 0.1$.

x	Euler	Midpoint	Exact
0.10	1.10000	1.11358	1.11517
0.20	1.22900	1.26072	1.26140
0.30	1.38790	1.43772	1.43986
0.40	1.57769	1.65026	1.65182
0.50	1.79946	1.89577	1.89872

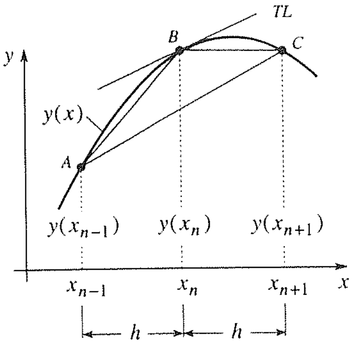


Figure 1. Graphical interpretation of midpoint rule versus Euler.

and

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_{n-1})}{2h}, \qquad \text{(midpoint)} \tag{14}$$

which are difference quotient approximations of the derivative $y'(x_n)$. In Fig. 1, we can interpret (14) and (13) as approximating $y'(x_n)$ by the slopes of the chords AC and BC, respectively, while the exact $y'(x_n)$ is the slope of the tangent line TL at x_n. We can see from the figure that AC gives a more accurate approximation than BC.

6.3.2. Second-order Runge–Kutta. The Runge–Kutta methods are developed somewhat differently. Observe that the low-order Euler method $y_{n+1} = y_n + f(x_n, y_n)h$ amounts to an extrapolation away from the initial point (x_n, y_n) using the slope $f(x_n, y_n)$ at that point. Expecting an average slope to give greater accuracy, one might try the algorithm

$$y_{n+1} = y_n + \frac{1}{2}\left[f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right] h, \tag{15}$$

which uses an average of the slopes at the initial and final points. Unfortunately, the formula (15) does not give y_{n+1} explicitly since y_{n+1} appears not only on the left-hand side but also in the argument of $f(x_{n+1}, y_{n+1})$. Intuition tells us that we should still do well if we replace that y_{n+1} by an estimated value, say, the Euler estimate $y_{n+1} = y_n + f(x_n, y_n)h$. Then the revised version of (15) is

$$y_{n+1} = y_n + \frac{1}{2}\left\{f(x_n, y_n) + f\left[x_{n+1}, y_n + f(x_n, y_n)h\right]\right\} h. \tag{16}$$

Thus, guided initially by intuition we can put the idea on a rigorous basis by considering a method of the form

$$y_{n+1} = y_n + \left\{a f(x_n, y_n) + b f\left[x_n + \alpha h, y_n + \beta f(x_n, y_n)h\right]\right\} h \tag{17}$$

and choosing the adjustable parameters a, b, α, β so as to make the order of the method (17) as high as possible; α, β determine the second slope location, and a, b

determine the "weights" of the two slopes. That is, we seek a, b, α, β so that the left- and right-hand sides of

$$\begin{aligned} y(x_{n+1}) \approx y(x_n) + \{af\left[x_n, y(x_n)\right] \\ +bf\left[x_n + \alpha h, y(x_n) + \beta f\left[x_n, y(x_n)\right] h\right]\} h \end{aligned} \tag{18}$$

agree to as high a degree in h as possible. Thus, expand the left-hand side (LHS) and right-hand side (RHS) of (18) in Taylor series in h:

$$\begin{aligned} \text{LHS} &= y(x_n) + y'(x_n)h + \frac{y''(x_n)}{2}h^2 + \cdots \\ &= y + fh + \frac{1}{2}\left(f_x + f_y f\right) h^2 + \cdots \end{aligned} \tag{19}$$

where y means $y(x_n)$ and the arguments of f, f_x, f_y are $x_n, y(x_n)$. Similarly (Exercise 9),

$$\text{RHS} = y + (a+b)fh + \left(\alpha f_x + \beta f f_y\right) bh^2 + \cdots. \tag{20}$$

Matching the h terms requires that

$$a + b = 1. \tag{21a}$$

Matching the h^2 terms, for any function f requires that

$$\alpha b = \frac{1}{2} \quad \text{and} \quad \beta b = \frac{1}{2}. \tag{21b}$$

The outcome then is that any method (17), with a, b, α, β chosen so as to satisfy (21), has a local truncation error that is $O(h^3)$ and is therefore a second-order method [subject to mild conditions on f such as the continuity of f and its first- and second-order partial derivatives so that we can justify the steps in deriving (19) and (20)]. These are the **Runge–Kutta methods of second order**.†

For instance, with $a = b = 1/2$ and $\alpha = \beta = 1$ we have

$$y_{n+1} = y_n + \frac{1}{2}\left\{f(x_n, y_n) + f\left[x_{n+1}, y_n + f(x_n, y_n)h\right]\right\} h, \tag{22}$$

which is usually expressed in the computationally convenient form

$$\boxed{\begin{aligned} &\qquad y_{n+1} = y_n + \tfrac{1}{2}\left(k_1 + k_2\right), \\ &k_1 = hf\left(x_n, y_n\right), \qquad k_2 = hf\left(x_{n+1}, y_n + k_1\right). \end{aligned}} \tag{23}$$

†The Runge–Kutta method was originated by *Carl D. Runge* (1856–1927), a German physicist and mathematician, and extended by the German aerodynamicist and mathematician *M. Wilhelm Kutta* (1867–1944). Kutta is well known for the Kutta–Joukowski formula for the lift on an airfoil, and for the "Kutta condition" of classical airfoil theory.

To understand this result, note that Euler's method would give $y_{n+1} = y_n + f(x_n, y_n)\, h$. If we denote that Euler estimate as y_{n+1}^{Euler}, then (22) can be expressed as

$$y_{n+1} = y_n + \frac{f(x_n, y_n) + f\left(x_{n+1}, y_{n+1}^{\text{Euler}}\right)}{2} h.$$

That is, we take a tentative step using Euler's method, then we average the slopes at the initial point x_n, y_n and at the Euler estimate of the final point $x_{n+1}, y_{n+1}^{\text{Euler}}$, and then make another Euler step, this time using the improved (average) slope. For this reason (23) is also known as the **improved Euler method**.

A different choice, $a = 0, b = 1, \alpha = \beta = 1/2$, gives what is known as the **modified Euler method**.

EXAMPLE 2. Let us proceed through the first two steps of the improved Euler method (23) for the same test problem as was used in Example 1,

$$y' = y + 2x - x^2; \qquad y(0) = 1, \qquad (0 \le x < \infty) \tag{24}$$

with $h = 0.1$; a more detailed illustration is given in Section 6.3.3 below. Here, $f(x, y) = y + 2x - x^2$.

$n = 0:$

$$\begin{aligned} k_1 &= hf(x_0, y_0) = 0.1\left[1 + 0 - (0)^2\right] = 0.1, \\ k_2 &= hf(x_1, y_0 + k_1) = 0.1\left[(1 + 0.1) + 2(0.1) - (0.1)^2\right] = 0.129, \\ y_1 &= y_0 + \tfrac{1}{2}(k_1 + k_2) = 1 + 0.5\,(0.1 + 0.129) = 1.1145; \end{aligned}$$

$n = 1:$

$$\begin{aligned} k_1 &= hf(x_1, y_1) = 0.1\left[1.1145 + 2(0.1) - (0.1)^2\right] = 0.13045, \\ k_2 &= hf(x_2, y_1 + k_1) \\ &= 0.1\left[(1.1145 + 0.13045) + 2(0.2) - (0.2)^2\right] = 0.160495, \\ y_2 &= y_1 + \tfrac{1}{2}(k_1 + k_2) = 1.1145 + 0.5\,(0.13045 + 0.160495) = 1.2600, \end{aligned}$$

compared with the values $y(x_1) = y(0.1) = 1.1152$ and $y(x_2) = y(0.2) = 1.2614$ obtained from the known exact solution $y(x) = x^2 + e^x$. ▮

6.3.3. Fourth-order Runge–Kutta. Using this idea of a weighted average of slopes at various points in the x, y plane, with the weights and locations determined so as to maximize the order of the method, one can derive higher-order Runge–Kutta methods as well, although the derivations are quite tedious. One of the most commonly used is the **fourth-order Runge–Kutta method**:

$$\boxed{\begin{array}{ll} \multicolumn{2}{c}{y_{n+1} = y_n + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)} \\ k_1 = hf\left(x_n, y_n\right), & k_2 = hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}k_1\right), \\ k_3 = hf\left(x_n + \frac{h}{2}, y_n + \frac{1}{2}k_2\right), & k_4 = hf\left(x_{n+1}, y_n + k_3\right), \end{array}} \tag{25}$$

which we give without derivation. Here the effective slope used is a weighted average of the slopes at the four points (x_n, y_n), $(x_n + h/2, y_n + k_1/2)$, $(x_n + h/2, y_n + k_2/2)$ and $(x_{n+1}, y_n + k_3)$ in the x, y plane, an average because the sum of the coefficients $1/6$, $2/6$, $2/6$, $1/6$ that multiply the k's is 1. Similarly, the sum of the coefficients $1/2, 1/2$ in the second-order version (23) is 1 as well.

EXAMPLE 3. As a summarizing illustration, we solve another "test problem,"

$$y' = -y; \qquad y(0) = 1 \tag{26}$$

by each of the methods considered, using a step size of $h = 0.05$ and single precision arithmetic (on the computer used that amounts to carrying eight significant figures; double precision would carry 16). The results are given in Table 2, together with the exact solution $y(x) = e^{-x}$ for comparison; $0.529\text{E}+2$, for instance, means 0.529×10^2. The value of y_1 for the midpoint rule was obtained by Euler's method with a reduced step size of 0.0025.

To illustrate the fourth-order Runge–Kutta calculation, let us proceed through the first step:

$n = 0:$

$$\begin{aligned} k_1 &= hf\left(x_0, y_0\right) = -hy_0 = -0.05(1) = -0.05, \\ k_2 &= hf\left(x_0 + \tfrac{h}{2}, y_0 + \tfrac{1}{2}k_1\right) = -h\left(y_0 + \tfrac{1}{2}k_1\right) \\ &= -0.05(1 - 0.025) = -0.04875, \\ k_3 &= hf\left(x_0 + \tfrac{h}{2}, y_0 + \tfrac{1}{2}k_2\right) = -h\left(y_0 + \tfrac{1}{2}k_2\right) \\ &= -0.05(1 - 0.024375) = -0.04878125, \\ k_4 &= hf\left(x_1, y_0 + k_3\right) = -h\left(y_0 + k_3\right) \\ &= -0.05(1 - 0.04878125) = -0.047560938, \end{aligned}$$

$$y_1 = y_0 + \tfrac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right) = 0.95122943,$$

which final result does agree with the corresponding entry in Table 2. Actually, there is a discrepancy of 2 in the last digit, but an error of that size is not unreasonable in view of the fact that the machine used carried only eight significant figures.

Most striking is the excellent accuracy of the fourth-order Runge–Kutta method, with six significant figure accuracy over the entire calculation.

COMMENT. We see that the midpoint rule and the second-order Runge–Kutta method yield comparable results initially, but the midpoint rule eventually develops an error that

Table 2. Comparison of the Euler, midpoint rule, second-order Runge–Kutta, fourth-order Runge–Kutta, and exact solutions of the initial-value problem (26), with $h = 0.05$.

x	Euler	Midpoint	2nd-order Runge–Kutta	4th-order Runge–Kutta	Exact $= e^{-x}$
0.00	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1	1.00000000 E+1
0.05	0.94999999 E+1	0.95116991 E+1	0.95125002 E+1	0.95122945 E+1	0.95122945 E+1
0.10	0.90249997 E+1	0.90488303 E+1	0.90487659 E+1	0.90483743 E+1	0.90483743 E+1
0.15	0.85737497 E+1	0.86068159 E+1	0.86076385 E+1	0.86070800 E+1	0.86070800 E+1
0.20	0.81450623 E+1	0.81881487 E+1	0.81880164 E+1	0.81873077 E+1	0.81873077 E+1
0.25	0.77378094 E+1	0.77880013 E+1	0.77888507 E+1	0.77880079 E+1	0.77880079 E+1
0.30	0.73509192 E+1	0.74093485 E+1	0.74091440 E+1	0.74081820 E+1	0.74081820 E+1
⋮					
2.00	0.12851217 E+0	0.13573508 E+0	0.13545239 E+0	0.13533530 E+0	0.13533528 E+0
2.05	0.12208656 E+0	0.12853225 E+0	0.12884909 E+0	0.12873492 E+0	0.12873492 E+0
2.10	0.11598223 E+0	0.12288185 E+0	0.12256770 E+0	0.12245644 E+0	0.12245644 E+0
2.15	0.11018312 E+0	0.11624406 E+0	0.11659253 E+0	0.11648417 E+0	0.11648415 E+0
2.20	0.10467397 E+0	0.11125745 E+0	0.11090864 E+0	0.11080316 E+0	0.11080315 E+0
2.25	0.99440269 E−1	0.10511832 E+0	0.10550185 E+0	0.10539923 E+0	0.10539922 E+0
2.30	0.94468258 E−1	0.10074562 E+0	0.10035863 E+0	0.10025885 E+0	0.10025885 E+0
⋮					
5.00	0.59205294 E−2	0.12618494 E−1	0.67525362 E−2	0.67379479 E−2	0.67379437 E−2
5.05	0.56245029 E−2	0.25511871 E−3	0.64233500 E−2	0.64093345 E−2	0.64093322 E−2
5.10	0.53432779 E−2	0.12592983 E−1	0.61102118 E−2	0.60967477 E−2	0.60967444 E−2
5.15	0.50761141 E−2	-0.10041796 E−2	0.58123390 E−2	0.57994057 E−2	0.57994043 E−2
5.20	0.48223082 E−2	0.12693400 E−1	0.55289874 E−2	0.55165654 E−2	0.55165626 E−2
5.25	0.45811930 E−2	-0.22735195 E−2	0.52594491 E−2	0.52475194 E−2	0.52475161 E−2
5.30	0.43521333 E−2	0.12920752 E−1	0.50030509 E−2	0.49915947 E−2	0.49915928 E−2
⋮					
9.70	0.47684727 E−4	0.64383668 E+0	0.61541170 E−4	0.61283507 E−4	0.61283448 E−4
9.75	0.45300490 E−4	-0.67670959 E+0	0.58541038 E−4	0.58294674 E−4	0.58294663 E−4
9.80	0.43035467 E−4	0.71150762 E+0	0.55687164 E−4	0.55451608 E−4	0.55451590 E−4
9.85	0.40883693 E−4	-0.74786037 E+0	0.52972413 E−4	0.52747200 E−4	0.52747171 E−4
9.90	0.38839509 E−4	0.78629363 E+0	0.50390008 E−4	0.50174691 E−4	0.50174654 E−4
9.95	0.36897534 E−4	-0.82648975 E+0	0.47933496 E−4	0.47727641 E−4	0.47727597 E−4
10.00	0.35052657 E−4	0.86894262 E+0	0.45596738 E−4	0.45399935 E−4	0.45399931 E−4

oscillates in sign, from step to step, and grows in magnitude. The reason for this strange (and incorrect) behavior will be studied in Section 6.5. ∎

Of course, in real applications we do not have the exact solution to compare with the numerical results. In that case, how do we know whether or not our results are sufficiently accurate? A useful rule of thumb, mentioned in Section 6.2, is to redo the entire calculation, each time with a smaller step size, until the results "settle down" to the desired number of significant digits.

Thus far we have taken h to be a constant, for simplicity, but there is no reason why it cannot be varied from one step to the next. In fact, there may be a compelling reason to do so. For instance, consider the equation $y' + y = \tanh 20x$ on $-10 \leq x \leq 10$. The function $\tanh 20x$ is almost a constant, except near the origin, where it varies dramatically approximately from -1 to $+1$. Thus, we need a very fine step size h near the origin for good accuracy, but to use that h over the entire x interval would be wasteful in terms of computer time and expense.

One can come up with a rational scheme for varying the step size to maintain a consistent level of accuracy, but such refinements are already available within existing software. For example, the default numerical differential equation solver in *Maple* is a "fourth-fifth order Runge–Kutta–Fehlberg method" denoted as RKF45 in the literature. According to RKF45, a tentative step is made, first using a fourth-order Runge–Kutta method, and then again using a fifth-order Runge–Kutta method. If the two results agree to a prespecified number of significant digits, then the fifth-order result is accepted. If they agree to more than that number of significant digits, then h is increased and the next step is made. If they agree to less than that number of significant digits, then h is decreased and the step is repeated.

6.3.4. Empirical estimate of the order. (Optional) The relative accuracies achieved by the different methods, as seen from the results in Table 2, strikingly reveal the importance of the order of the method. Thus, it is important to know how to verify the order of whatever method we use, if only as a partial check on the programming.

Recall that by a method being of order p we mean that at any chosen x the error behaves as Ch^p for some constant C:

$$y_{\text{exact}} - y_{\text{comp}} \sim Ch^p \tag{27}$$

as $h \to 0$. Suppose we wish to check the order of a given method. Select a test problem such as the one in Example 3, and use the method to compute y at any x point such as $x = 1$, for two different h's say h_1 and h_2. Letting $y_{\text{comp}}^{(1)}$ and $y_{\text{comp}}^{(2)}$ denote the y's computed at $x = 1$ using step sizes of h_1 and h_2, respectively, we have

$$y_{\text{exact}} - y_{\text{comp}}^{(1)} \approx Ch_1^p,$$
$$y_{\text{exact}} - y_{\text{comp}}^{(2)} \approx Ch_2^p.$$

Dividing one equation by the other, to cancel the unknown C, and solving for p, gives

$$p \approx \frac{\ln\left[\dfrac{y_{\text{exact}}-y_{\text{comp}}^{(1)}}{y_{\text{exact}}-y_{\text{comp}}^{(2)}}\right]}{\ln\left[\frac{h_1}{h_2}\right]}. \tag{28}$$

To illustrate, let us run the test problem (26) with Euler's method, with $h_1 = 0.1$ and $h_2 = 0.05$. The results at $x = 1$ are

$$\begin{aligned} h_1 &= 0.1, & y_{\text{comp}}^{(1)} &= 0.348678440100, \\ h_2 &= 0.05, & y_{\text{comp}}^{(2)} &= 0.358485922409, \end{aligned}$$

and since $y_{\text{exact}}(1) = 0.367879441171$, (28) gives $p \approx 1.03$, which is respectably close to 1. We should be able to obtain a more accurate estimate of p by using smaller h's since (27) holds only as $h \to 0$. In fact, using $h_1 = 0.05$ and $h_2 = 0.02$ gives $p \approx 1.01$. Using those same step sizes, we also obtain $p \approx 2.05, 2.02$, and 4.03 for the midpoint rule, second-order Runge–Kutta, and fourth-order Runge–Kutta methods, respectively.

Why not use even smaller h's to determine p more accurately? Two difficulties arise. One is that as the h's are decreased the computed solutions become more and more accurate, and the $y_{\text{exact}} - y_{\text{comp}}^{(1)}$ and $y_{\text{exact}} - y_{\text{comp}}^{(2)}$ differences in (28) are known to fewer and fewer significant figures, due to cancelation. This is especially true for a high-order method. The other difficulty is that (27) applies to the truncation error alone so, implicit in our use of (27) is the assumption that roundoff errors are negligible. If we make h too small, that assumption may become invalid. For both of these reasons it is important to use extended precision for such calculations, as we have for the preceding calculations.

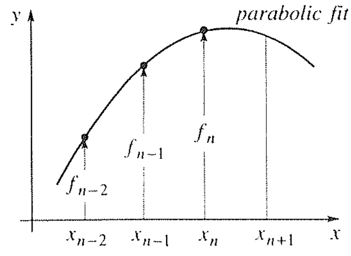


Figure 2. Adams–Bashforth interpolation of f for the case $m = 2$.

6.3.5. Multi-step and predictor-corrector methods. (Optional) We've already called attention to the multi-step nature of the midpoint rule. Our purpose in this optional section is to give a brief overview of a class of multi-step methods known as **Adams–Bashforth methods**, obtained by integrating $y' = f(x, y)$ from x_n to x_{n+1}:

$$\int_{x_n}^{x_{n+1}} y'\,dx = \int_{x_n}^{x_{n+1}} f(x, y(x))\,dx \tag{29}$$

or

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x))\,dx. \tag{30}$$

To evaluate the integral, we fit $f(x, y(x))$ with a polynomial of degree m, which is readily integrated. The polynomial interpolates (i.e., takes on the same values as) $f(x, y(x))$ at the $m + 1$ points $x_{n-m}, \dots, x_{n-1}, x_n$ as illustrated in Fig. 2 for the case $m = 2$. As the simplest case, let $m = 0$. Then the zeroth degree polynomial approximation of $f(x, y(x))$ on $[x_n, x_{n+1}]$ is $f(x, y(x)) \approx f_n$, where f_n denotes $f(x_n, y_n)$, and (30) gives the familiar Euler method $y_{n+1} = y_n + f_n h$. Omitting the steps in this overview we state that with $m = 3$ one obtains the fourth-order **Adams–Bashforth method**

$$y_{n+1} = y_n + (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3})\,\frac{h}{24} \tag{31a}$$

with a local truncation error of

$$(e_n)_{AB} = \frac{251}{720}\,y^{(v)}(\xi)h^5 \tag{31b}$$

for some ξ in $[x_{n-3}, x_n]$.We can see that (31a) is not self-starting; it applies only for $n = 3, 4, \ldots$, so the first three steps (for $n = 0, 1, 2$) need to be carried out by some other method.

Suppose that instead of interpolating f at $x_{n-m}, \ldots, x_{n-1}, x_n$ we interpolate at $x_{n-m+1}, \ldots, x_n, x_{n+1}$. With $m = 3$, again, one obtains the fourth-order **Adams–Moulton method**

$$y_{n+1} = y_n + (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2})\,\frac{h}{24} \tag{32a}$$

with a local truncation error of

$$(e_n)_{AM} = -\frac{19}{720}\,y^{(v)}(\xi)h^5, \tag{32b}$$

where the ξ's in (31b) and (32b) are different, in general.

Although both methods are fourth order, the Adams–Moulton method is more accurate because the constant factor in $(e_n)_{AM}$ is roughly thirteen times smaller than the constant in $(e_n)_{AB}$. This increase in accuracy occurs because the points $x_{n-m+1}, \ldots, x_n, x_{n+1}$ are more centered on the interval of integration (from x_n to x_{n+1}) than the points $x_{n-m}, \ldots, x_{n-1}, x_n$. On the other hand, the term $f_{n+1} = f(x_{n+1}, y_{n+1})$ in (32a) is awkward because the argument y_{n+1} is not yet known! [If f is linear in y, then (32a) can be solved for y_{n+1} by simple algebra, and the awkwardness disappears.] Thus, the method (32a) is said to be of **closed** type, whereas (31a) and all of our preceding methods have been of **open** type.

To handle this difficulty, it is standard practice to solve closed formulas by iteration. Using superscripts to indicate the iterate, (32a) becomes

$$y_{n+1}^{(k+1)} = y_n + \left[9f\left(x_{n+1}, y_{n+1}^{(k)}\right) + 19f_n - 5f_{n-1} + f_{n-2}\right]\frac{h}{24}. \tag{33}$$

To start the iteration, we compute $y_{n+1}^{(0)}$ from a **predictor formula**, with subsequent corrections made by the **corrector formula** (33). It is recommended that the predictor and corrector formulas be of the same order (certainly, the corrector should never be of lower order than the predictor) with the corrector applied only once. Thus, the Adams–Bashforth and Adams–Moulton methods constitute a natural **predictor-corrector** pair with "AB" as the predictor and "AM" as the corrector. Why might we choose the fourth-order AB–AM predictor-corrector over the Runge–Kutta method of the same order or vice versa? On the negative side, AB–AM is not self-starting, it requires the storage of f_{n-3}, f_{n-2}, and f_{n-1}, and is more tedious to program. On the other hand, it involves only two function evaluations per step (namely, f_n and f_{n+1}) if the corrector is applied only once, whereas Runge–Kutta involves four. Thus, if $f(x, y)$ is reasonably complicated then we can expect AB–AM to be almost twice as fast. In large-scale computing, the savings can be significant.

Closure. Motivated to seek higher-order methods than the first-order Euler method, we use a Taylor series approach to obtain the second-order midpoint rule. Though more accurate, a disadvantage of the midpoint rule is that it is not self-starting. Pursuing a different approach, we look at the possibility of using a weighted average of slopes at various points in the x, y plane, with the weights and locations determined so as to maximize the order of the method. We thereby derive the second-order Runge–Kutta method and present, without derivation, the fourth-order Runge–Kutta method. The latter is widely used because it is accurate and self-starting.

Because of the importance of the order of a given method, we suggest that the order be checked empirically using a test problem with a known exact solution. The resulting approximate expression for the order is given by (28).

In the final section we return to the idea of multistep methods and present a brief overview of the Adams–Bashforth methods, derived most naturally from an approximate integral approach. Though not self-starting, the fourth-order Adams-Bashforth method (31a) is faster than the Runge–Kutta method of the same order because it requires only one function evaluation per step (namely, f_n; the f_{n-1}, f_{n-2}, and f_{n-3} terms are stored from previous steps). A further refinement consists of predictor-corrector variations of the Adams–Bashforth methods. However, we stress that such refinements become worthwhile only if the scope of the computational effort becomes large enough to justify the additional inconvenience caused by such features as the absence of self-starting and predictor-corrector iteration. Otherwise, one might as well stick to a simple and accurate method such as fourth-order Runge–Kutta.

Computer software. Computer-software systems such as *Maple* include numerical differential equation solvers. In *Maple* one can use the dsolve command together with a numeric option. The default numerical solution method is the RKF45 method mentioned above. Note that with the numeric option of dsolve one does not specify a step size h since that choice is controlled within the program and, in general, is varied from step to step to maintain a certain level of accuracy. To specify the absolute error tolerance one can use an additional option called abserr, which is formatted as abserr = Float(1,2-digits) and which means 1 times 10 to the one- or two-digit exponent. For instance, to solve

$$y' = -y; \quad y(0) = 1$$

for $y(x)$ with an absolute error tolerance of 1×10^{-5}, and to print the results at $x = 2, 10$, enter

```
with(DEtools):
```

and return. Then enter

dsolve({diff($y(x), x) = -y(x)$, $y(0) = 1$}, $y(x)$, type = numeric,
value = array([2, 10]), abserr = Float(1, −5));

and return. The printed result is

$$\begin{bmatrix} [x, y(x)] \\ \begin{bmatrix} 2. & .1353337989380555 \\ 10. & .00004501989255717160 \end{bmatrix} \end{bmatrix}$$

For comparison, the exact solution is $y(2) = \exp(-2) = 0.1353352832$ and $y(10) = \exp(-10) = 0.0000453999$, respectively.

EXERCISES 6.3

1. Evaluate y_1 and y_2 by hand, by the second-order and fourth-order Runge–Kutta methods, with $h = 0.02$. Obtain the exact values $y(x_1)$ and $y(x_2)$ as well.

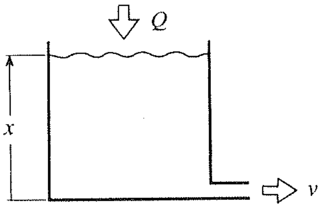
(a) $y' = 3000xy^{-2}$; $y(0) = 2$
(b) $y' = 40xe^{-y}$; $y(0) = 3$
(c) $y' = x + y$; $y(-1) = 5$
(d) $y' = -y \tan x$; $y(1) = -8$
(e) $y' = (y^2 + 1)/4$; $y(0) = 0$
(f) $y' = -2y \sin x$; $y(2) = 5$

2. (a)–(f) Program the second- and fourth-order Runge–Kutta methods and use them to solve the initial-value problem given in the corresponding part of Exercise 1 but with the initial condition $y(0) = 1$. Use $h = 0.05$. Print out all computed values of y, up to $x = 0.5$, as well as the exact solution.

3. Using the test problem (10), do an empirical evaluation of the order of the given method. Use (28), with $h = 0.1$ and 0.05, say. Do the evaluation at two different locations, such as $x = 1$ and $x = 2$. (The order should not depend upon x so your results at the two points should be almost identical.)

(a) Euler's method
(b) Second-order Runge–Kutta method
(c) Fourth-order Runge–Kutta method

4. (*Liquid level*) Liquid is pumped into a tank of horizontal cross-sectional area A (m^2) at a rate Q (liters/sec), and is drained by a valve at its base as sketched in the figure.



According to Bernoulli's principle, the efflux velocity $v(t)$ is approximately $\sqrt{2gx(t)}$, where g is the acceleration of gravity. Thus, a mass balance gives

$$\begin{aligned} Ax'(t) &= Q(t) - Bv(t) \\ &= Q(t) - B\sqrt{2gx(t)}, \end{aligned} \tag{4.1}$$

where B is the cross-sectional area of the efflux pipe. For definiteness, suppose that $A = 1$ and $B\sqrt{2g} = 0.01$ so

$$x' = Q(t) - 0.01\sqrt{x}. \tag{4.2}$$

We wish to know the depth $x(t)$ at the end of 10 minutes ($t = 600$ sec), 20 minutes, . . . , up to one hour. Program the computer solution of (4.2) by the second-order Runge–Kutta method for the following cases, and use it to solve for those x values: $x(600), x(1200), \ldots, x(3600)$. (Using the rule of thumb given below Example 3, reduce h until those results settle down to four significant digits.)

(a) $Q(t) = 0.02$; $x(0) = 0$
(b) $Q(t) = 0.02$; $x(0) = 2$
(c) $Q(t) = 0.02$; $x(0) = 4$
(d) $Q(t) = 0.02$; $x(0) = 6$

(e) $Q(t) = 0.02\left(1 - e^{-0.004t}\right); \quad x(0) = 0$
(f) $Q(t) = 0.02\left(1 - e^{-0.004t}\right); \quad x(0) = 8$
(g) $Q(t) = 0.02t; \quad x(0) = 0$
(h) $Q(t) = 0.02(1 + \sin 0.1t); \quad x(0) = 0$
NOTE: Surely, we will need h to be small compared to the period 200π of $Q(t)$ in part (h).

5. (a)–(h) (*Liquid level*) Same as Exercise 4, but use fourth-order Runge–Kutta instead of second order.

6. (a)–(h) (*Liquid level*) Same as Exercise 4, but use computer software to do the numerical solution of the differential equation. In *Maple*, for instance, the dsolve command uses the fourth-fifth order RKF45 method.

7. (*Liquid level*) (a) For the case where $Q(t)$ is a constant, derive the general solution of (4.2) in Exercise 4 as

$$Q - 0.01\sqrt{x} - Q\ln\left(Q - 0.01\sqrt{x}\right) = 0.00005t + C, \quad (7.1)$$

where C is the constant of integration.
(b) Evaluate C in (7.1) if $Q = 0.02$ and $x(0) = 0$. Then, solve (7.1) for $x(t)$ at $t = 600, 1200, \dots, 3600$. NOTE: Unfortunately, (7.1) is in implicit rather than explicit form, but you can use computer software to solve. In *Maple*, for instance, the relevant command is fsolve.

8. Suppose that we have a convergent method, with $E_n \sim Ch^p$ as $h \to 0$. Someone offers to improve the method by either halving C or by doubling p. Which would you choose? Explain.

9. Expand the right-hand side of (18) in a Taylor series in h and show that the result is as given in (20). HINT: To expand the $f\left(x_n + \alpha h, y + \beta f h\right)$ term you need to use chain differentiation.

10. (a) Program the fourth-order Runge–Kutta method (25) and use it to run the test problem (10) and to compute y at $x = 1$ using $h = 0.05$ and then $h = 0.02$. From those values and the known exact solution, empirically verify that the method is fourth order.
(b) To see what harm a programming error can cause, change the $x_n + h/2$ in the formula for k_2 to x_n, repeat the two evaluations of y at $x = 1$ using $h = 0.05$ and $h = 0.02$, and empirically determine the order of the method. Is it still a fourth-order method?
(c) Instead of introducing the programming error suggested in part (b), suppose we change the coefficient of k_2 in $y_{n+1} = y_n + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)$ from 2 to 3. Do you think the method will still be convergent? Explain.

11. (*Rectangular, trapezoidal, and Simpson's rule*) Consider the special case where f in $y' = f$ is a function of x only. Integrating $y' = f(x)$ from x_n to x_{n+1}, we have

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x)\,dx \qquad (11.1)$$

If we fit $f(x)$, over $[x_n, x_{n+1}]$, with a *zeroth*-degree polynomial (i.e., a constant) that interpolates f at x_n, then we have $f(x) \approx f(x_n)$, and (11.1) gives $y(x_{n+1}) \approx y(x_n) + f(x_n)h$ and hence the Euler method $y_{n+1} = y_n + f(x_n)h$.

(a) Show that if we fit $f(x)$, over $[x_n, x_{n+1}]$, with a *first*-degree polynomial (a straight line) that interpolates f at x_n and x_{n+1}, then $f(x) \approx f(x_n) + \left[f(x_{n+1}) - f(x_n)\right](x - x_n)/h$. Putting that approximation into (11.1), derive the approximation

$$y(x_{n+1}) = y(x_n) + \frac{1}{2}\left[f(x_n) + f(x_{n+1})\right]h, \qquad (11.2)$$

and show that (for the special case where f is a function of x only) (11.2) is identical to the second-order Runge–Kutta method (23).
(b) Show that if we fit $f(x)$, over $[x_n, x_{n+1}]$, with a second-degree polynomial (a parabola) that interpolates f at x_n, $x_n + h/2$, and $x_{n+1} = x_n + h$, and put that approximation into (11.1), then one obtains

$$y(x_{n+1}) = y(x_n) + \frac{1}{6}\left[f(x_n) + 4f(x_n + h/2) + f(x_{n+1})\right]h, \qquad (11.3)$$

and show that (for the case where f is a function of x only) (11.3) is identical to the fourth-order Runge–Kutta method (25). NOTE: These three results amount to the well-known **rectangular**, **trapezoidal**, and **Simpson's** rules of numerical integration for a single interval of size h. If we sum over all of the intervals, they take the forms

$$\int_a^b f(x)\,dx = \begin{cases} \left[f(a) + f(a+h) + \cdots + f(b)\right]h, \\ \left[f(a) + 2f(a+h) + 2f(a+2h) \right. \\ \qquad \left. + \cdots + 2f(b-h) + f(b)\right]\frac{h}{2}, \\ \left[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) \right. \\ \qquad \left. + \cdots + 4f(b-h) + f(b)\right]\frac{h}{6}; \end{cases} \qquad (11.4)$$

respectively. In passing from (11.3) to the last line of (11.4) we have replaced $h/2$ by h everywhere in

(11.3), so that the revised (11.3) reads $y_{n+1} = y_n + [f(x_n) + 4f(x_{n+1}) + f(x_{n+2})]\, h/3$, where $x_n = a+nh$. For the rectangular and trapezoidal rules the number of subdivisions, $(b-a)/h$, can be even or odd, but for Simpson's rule it must be even. The order of the error for these integration methods is $O(h)$, $O(h^2)$, and $O(h^4)$, respectively.

12. (a) Using $m = 1$, derive from (30) the Adams–Bashforth method

$$y_{n+1} = y_n + (3f_n - f_{n-1})\,\frac{h}{2}. \tag{12.1}$$

(b) Determine the order of the method (12.1) empirically by using it to solve the test problem (10), at $x = 1$, with two different step sizes, and then using (28).

13. This exercise is to take you through the fourth-order AB–AM predictor-corrector scheme.
(a) For the problem $y' = 2xy^2, y(0) = 1$, compute y_1, y_2, y_3 from the exact solution, with $h = 0.1$, and use those as starting values to determine y_4, by hand, by means of the fourth-order AB–AM predictor-corrector scheme given by (31a) and (33). Apply the corrector three times.
(b) Continuing in the same way, determine y_5.

6.4 Application to Systems and Boundary-Value Problems

The methods developed in the preceding sections are for an initial-value problem with a single first-order differential equation, but what if we have a system of differential equations, a higher-order equation, or a boundary-value problem? In this section we extend the Euler and fourth-order Runge–Kutta methods to these cases.

6.4.1. Systems and higher-order equations. Consider the system of initial-value problems

$$u'(x) = f(x, u, v); \qquad u(a) = u_0 \tag{1a}$$

$$v'(x) = g(x, u, v); \qquad v(a) = v_0 \tag{1b}$$

on $u(x)$ and $v(x)$. To extend Euler's method to such a system, we merely apply it to each of the problems (1a) and (1b) as follows:

$$\begin{aligned} u_{n+1} &= u_n + f(x_n, u_n, v_n)h, \\ v_{n+1} &= v_n + g(x_n, u_n, v_n)h \end{aligned} \tag{2}$$

for $n = 0, 1, 2, \ldots$. Equations (2) are coupled (since each involves both u and v), as were equations (1), but that coupling causes no complication because it occurs in the right-hand sides of the equations, and the values u_n, v_n are already known from the preceding step.

EXAMPLE 1. Consider the system

$$\begin{aligned} u' &= x + v; \qquad & u(0) &= 0 \\ v' &= uv^2; \qquad & v(0) &= 1. \end{aligned} \tag{3}$$

The latter looks fairly simple, but it is not. It is nonlinear because of the uv^2 term. Turning to numerical solution using the Euler method (2), let $h = 0.1$, say, and let us go through the first couple of steps. First, $u_0 = 0$ and $v_0 = 1$ from the initial conditions. Then,

$$\begin{aligned}
n = 0:&\\
&u_1 = u_0 + (x_0 + v_0)h = 0 + (0 + 1)(0.1) = \underline{0.1},\\
&v_1 = v_0 + u_0 v_0^2 h = 1 + (0)(1)^2(0.1) = \underline{1}.\\
n = 1:&\\
&u_2 = u_1 + (x_1 + v_1)h = 0.1 + (0.1 + 1)(0.1) = \underline{0.21},\\
&v_2 = v_1 + u_1 v_1^2 h = 1 + (0.1)(1)^2(0.1) = \underline{1.01},
\end{aligned}$$

and so on. ▮

Similarly, if the system contains more than two equations.

Next, we show how to adapt the fourth-order Runge–Kutta method to the system (1). Recall that for the single equation

$$y' = f(x, y); \qquad y(a) = y_0 \tag{4}$$

the algorithm is

$$\begin{aligned}
&y_{n+1} = y_n + \tfrac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right),\\
&k_1 = hf(x_n, y_n), \qquad k_2 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_1\right),\\
&k_3 = hf\left(x_n + \tfrac{h}{2}, y_n + \tfrac{1}{2}k_2\right), \quad k_4 = hf\left(x_{n+1}, y_n + k_3\right).
\end{aligned} \tag{5}$$

For the system (1) it becomes

$$\begin{aligned}
u_{n+1} &= u_n + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right),\\
v_{n+1} &= v_n + \frac{1}{6}\left(l_1 + 2l_2 + 2l_3 + l_4\right),
\end{aligned}$$

$$\begin{aligned}
k_1 &= hf(x_n, u_n, v_n),\\
l_1 &= hg(x_n, u_n, v_n),\\
k_2 &= hf\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_1, v_n + \frac{1}{2}l_1\right),\\
l_2 &= hg\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_1, v_n + \frac{1}{2}l_1\right),\\
k_3 &= hf\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_2, v_n + \frac{1}{2}l_2\right),
\end{aligned} \tag{6}$$

$$\begin{aligned}
l_3 &= hg\left(x_n + \frac{h}{2}, u_n + \frac{1}{2}k_2, v_n + \frac{1}{2}l_2\right),\\
k_4 &= hf\left(x_{n+1}, u_n + k_3, v_n + l_3\right),\\
l_4 &= hg\left(x_{n+1}, u_n + k_3, v_n + l_3\right),
\end{aligned}$$

and similarly for systems containing more than two equations.

EXAMPLE 2. Let us illustrate (6) using the same system as in Example 1,

$$\begin{aligned}
u' &= x + v; \qquad & u(0) &= 0\\
v' &= uv^2; \qquad & v(0) &= 1.
\end{aligned} \tag{7}$$

With $h = 0.1$, say, (6) gives

$n = 0:$

$$\begin{aligned}
k_1 &= h(x_0 + v_0) = (0.1)(0 + 1) = 0.1,\\
l_1 &= hu_0v_0^2 = (0.1)(0)(1)^2 = 0,\\
k_2 &= h\left[\left(x_0 + \frac{h}{2}\right) + \left(v_0 + \frac{l_1}{2}\right)\right] = (0.1)\left[(0 + 0.05) + (1 + 0)\right] = 0.105,\\
l_2 &= h\left(u_0 + \frac{k_1}{2}\right)\left(v_0 + \frac{l_1}{2}\right)^2 = (0.1)\left(0 + 0.05\right)\left(1 + 0\right)^2 = 0.005,\\
k_3 &= h\left[\left(x_0 + \frac{h}{2}\right) + \left(v_0 + \frac{l_2}{2}\right)\right]\\
&= (0.1)\left[(0 + 0.05) + (1 + 0.0025)\right] = 0.10525,\\
l_3 &= h\left(u_0 + \frac{k_2}{2}\right)\left(v_0 + \frac{l_2}{2}\right)^2\\
&= (0.1)\left(0 + 0.0525\right)\left(1 + 0.0025\right)^2 = 0.005276,\\
k_4 &= h\left[x_1 + (v_0 + l_3)\right]\\
&= (0.1)\left[0.1 + (1 + 0.005276)\right] = 0.110528,\\
l_4 &= h\left(u_0 + k_3\right)\left(v_0 + l_3\right)^2\\
&= (0.1)\left(0 + 0.10525\right)\left(1 + 0.005276\right)^2 = 0.010636,\\
u_1 &= u_0 + \frac{1}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right)\\
&= 0 + \frac{1}{6}\left(0.1 + 0.21 + 0.2105 + 0.110528\right) = \underline{0.105171},\\
v_1 &= v_0 + \frac{1}{6}\left(l_1 + 2l_2 + 2l_3 + l_4\right)\\
&= 1 + \frac{1}{6}\left(0 + 0.01 + 0.010552 + 0.010636\right) = \underline{1.005198}.
\end{aligned}$$

$$
\begin{aligned}
n = 1: \quad & \\
& k_1 = 0.110520, \qquad l_1 = 0.010627, \\
& k_2 = 0.116051, \qquad l_2 = 0.016382, \\
& k_3 = 0.116339, \qquad l_3 = 0.016760, \\
& k_4 = 0.122196, \qquad l_4 = 0.023134, \\
& u_2 = \underline{0.221420}, \qquad v_2 = \underline{1.021872},
\end{aligned}
$$

and so on for $n = 2, 3, \ldots$. We suggest that you fill in the details for the calculation of the $k_1, \ldots, v_2$ values shown above for $n = 1$. ▮

Of course, the idea is to carry out such calculations on a computer, not by hand. The calculations shown in Examples 1 and 2 are merely intended to clarify the methods.

What about higher-order equations? The key is to re-express an nth-order equation as an equivalent system of n first-order equations.

EXAMPLE 3. The problem

$$y''' - xy'' + y' - 2y^3 = \sin x; \qquad y(1) = 2,\ y'(1) = 0,\ y''(1) = -3 \tag{8}$$

can be converted to an equivalent system of three first-order equations as follows. Define $y' = u$ and $y'' = v$ (hence $u' = v$). Then (8) can be re-expressed in the form

$$
\begin{aligned}
y' &= u; & y(1) &= 2 \\
u' &= v; & u(1) &= 0 \\
v' &= \sin x + 2y^3 - u + xv; & v(1) &= -3.
\end{aligned}
\tag{9a,b,c}
$$

Of the three differential equations in (9), the first two merely serve to introduce the auxiliary dependent variables u and v, and since v' is y''' the third one is a restated version of the given equation $y''' - xy'' + y' - 2y^3 = \sin x$. Equation (9a) is the y equation, so the initial condition is on $y(1)$, namely, $y(1) = 2$, as given in (8). Equation (9b) is the u equation, so the initial condition is on $u(1)$, and we have $u(1) = y'(1) = 0$, from (8). Similarly, for equation (9c).

The system (9) can now be solved by the Euler or fourth-order Runge–Kutta methods or any other such algorithm. To illustrate, let us carry out the first two steps using Euler's method, taking $h = 0.2$, say.

$$
\begin{aligned}
n = 0: \quad & \\
& y_1 = y_0 + u_0 h = 2 + (0)(0.2) = \underline{2}, \\
& u_1 = u_0 + v_0 h = 0 + (-3)(0.2) = -0.6, \\
& v_1 = v_0 + \left(\sin x_0 + 2y_0^3 - u_0 + x_0 v_0\right) h \\
& \quad = -3 + \left[\sin 1 + 2(2)^3 - 0 + (1)(-3)\right](0.2) = -0.231706. \\
n = 1: \quad & \\
& y_2 = y_1 + u_1 h = 2 + (-0.6)(0.2) = \underline{1.88},
\end{aligned}
$$

$$\begin{aligned}
u_2 &= u_1 + v_1 h = -0.6 + (-0.231706)(0.2) = -0.646341, \\
v_2 &= v_1 + \left(\sin x_1 + 2y_1^3 - u_1 + x_1 v_1\right) h \\
&= -0.231706 + \left[\sin 0.2 + 2(2)^3 - (-0.6) + (0.2)(-0.231706)\right](0.2) \\
&= 3.118760,
\end{aligned}$$

and so on for $n = 2, 3, \ldots$.

COMMENT. Observe that at each step we compute y, u, and v, yet we are not really interested in the auxiliary variables u and v. Perhaps we could just compute $y_1, y_2, \ldots$ and not the u, v values? No; equations (9) are coupled so we need to bring all three variables along together. Of course, we don't need to print or plot u and v, but we do need to compute them. ▮

EXAMPLE 4. Examples 1 and 2 involve a system of first-order equations, and Example 3 involve a single higher-order equation. As a final example, consider a combination of the two such as the initial-value problem

$$\begin{aligned}
&u'' - 3xuv = \sin x; \quad u(0) = 4, \quad u'(0) = -1 \\
&v'' + 2u - v = 5x; \quad v(0) = 7, \quad v'(0) = 0.
\end{aligned} \tag{10}$$

The idea is exactly the same as before. We need to recast (10) as a system of first-order initial value problems. We can do so by introducing auxiliary dependent variables w and z according to $u' = w$ and $v' = z$. Then (10) becomes

$$\begin{aligned}
u' &= w; & u(0) &= 4 \\
w' &= \sin x + 3xuv; & w(0) &= -1 \\
v' &= z; & v(0) &= 7 \\
z' &= 5x - 2u + v; & z(0) &= 0
\end{aligned} \tag{11}$$

which system can now be solved by Euler's method or any other such numerical differential equation solver. ▮

6.4.2. Linear boundary-value problems. Our discussion is based mostly upon the following example.

EXAMPLE 5. Consider the third-order boundary-value problem

$$y''' - x^2 y = -x^4; \qquad y(0) = 0, \ y'(0) = 0, \ y(2) = 4. \tag{12}$$

To solve numerically, we begin by recasting (12) as the first-order system:

$$\begin{aligned}
y' &= u; & y(0) &= 0, \quad y(2) = 4 \\
u' &= v; & u(0) &= 0 \\
v' &= x^2 y - x^4.
\end{aligned} \tag{13a,b,c}$$

However, we cannot apply the numerical integration techniques that we have discussed because the problem (13c) does not have an initial condition so we cannot get the solution started. Whereas (13c) is missing an initial condition on v, (13a) has an extra condition – the right end condition $y(2) = 4$, but that condition is of no help in developing a numerical integration scheme that develops a solution beginning at $x = 0$.

Nevertheless, the linearity of (12) saves the day and permits us to work with an initial-value version instead. Specifically, suppose that we solve (numerically) the four initial-value problems

$$\begin{aligned} &L[Y_1] = 0, && Y_1(0) = 1, && Y_1'(0) = 0, && Y_1''(0) = 0, \\ &L[Y_2] = 0, && Y_2(0) = 0, && Y_2'(0) = 1, && Y_2''(0) = 0, \\ &L[Y_3] = 0, && Y_3(0) = 0, && Y_3'(0) = 0, && Y_3''(0) = 1, \\ &L[Y_p] = -x^4, && Y_p(0) = 0, && Y_p'(0) = 0, && Y_p''(0) = 0, \end{aligned} \tag{14}$$

where $L = d^3/dx^3 - x^2$ is the differential operator in (12). The nine initial conditions in the first three of these problems were chosen so as to have a nonzero determinant so that Y_1, Y_2, Y_3 comprise a fundamental set of solutions (i.e., a linearly independent set of solutions) of the homogeneous equation $L[Y] = 0$. The three initial conditions on the particular solution Y_p were chosen as zero for simplicity; any values will do since any particular solution will do. Suppose we imagine that the four initial-value problems in (14) have now been solved by the methods discussed above. Then Y_1, Y_2, Y_3, Y_p are known functions of x over the interval of interest $[0, 2]$, and we have the general solution

$$y(x) = C_1Y_1(x) + C_2Y_2(x) + C_3Y_3(x) + Y_p(x) \tag{15}$$

of $L[y] = -x^4$. Finally, we evaluate the integration constants C_1, C_2, C_3 by imposing the boundary conditions given in (12):

$$\begin{aligned} y(0) &= 0 = C_1 + 0 + 0 + 0, \\ y'(0) &= 0 = 0 + C_2 + 0 + 0, \\ y(2) &= 4 = C_1Y_1(2) + C_2Y_2(2) + C_3Y_3(2) + Y_p(2). \end{aligned} \tag{16}$$

Solving (16) gives $C_1 = C_2 = 0$ and $C_3 = [4 - Y_p(2)]/Y_3(2)$, so we have the desired solution of (12) as

$$y(x) = \frac{4 - Y_p(2)}{Y_3(2)} Y_3(x) + Y_p(x). \tag{17}$$

In fact, since $C_1 = C_2 = 0$ the functions $Y_1(x)$ and $Y_2(x)$, have dropped out so we don't need to calculate them. All we need are $Y_3(x)$ and $Y_p(x)$, and these are found by the numerical integration of the initial-value problems

$$\begin{aligned} Y_3' &= U_3, && Y_3(0) = 0, \\ U_3' &= V_3, && U_3(0) = 0, \\ V_3' &= x^2Y_3, && V_3(0) = 1, \end{aligned} \tag{18}$$

and

$$\begin{aligned} Y_p' &= U_p, & Y_p(0) &= 0, \\ U_p' &= V_p, & U_p(0) &= 0, \\ V_p' &= x^2 Y_p - x^4, & V_p(0) &= 0, \end{aligned} \tag{19}$$

respectively.

COMMENT. Remember that whereas initial-value problems have unique solutions (if the functions involved are sufficiently well behaved), boundary-value problems can have no solution, a unique solution, or even an infinite number of solutions. How do these possibilities work out in this example? The clue is that (17) fails if $Y_3(2)$ turns out to be zero. The situation is seen more clearly from (16), where all of the possibilities come into view. Specifically, if $Y_3(2) \neq 0$, then we can solve uniquely for C_3, and we have a unique solution, given by (17). If $Y_3(2)$ does vanish, then there are two possiblities as seen from (16): if $Y_p(2) \neq 4$, then there is no solution, and if $Y_p(2) = 4$ then there are an infinite number of solutions of (12), namely,

$$y(x) = C_3 Y_3(x) + Y_p(x), \tag{20}$$

where C_3 remains arbitrary. ▮

We see that boundary-value problems are more difficult than initial-value problems. From Example 5 we see that a nonhomogeneous nth-order linear boundary-value problem generally involves the solution of $n+1$ initial-value problems, although in Example 5 (in which $n = 3$) we were lucky and did not need to solve for two of the four unknowns, Y_1 and Y_2.

Nonlinear boundary-value problems are more difficult still, because we cannot use the idea of finding a fundamental set of solutions plus a particular solution and thus forming a general solution, as we did in Example 5, and which idea is based upon linearity. One viable line of approach comes under the heading of **shooting methods**. For instance, to solve the nonlinear boundary-value problem

$$y'' + \sin y = 3x; \qquad y(0) = 0, \;\; y(5) = 2 \tag{21}$$

we can solve the initial-value problem

$$\begin{aligned} y' &= u, & y(0) &= 0 \\ u' &= 3x - \sin y, & u(0) &= u_0 \end{aligned} \tag{22}$$

iteratively. That is, we can guess at the initial condition u_0 [which is the initial slope $y'(0)$] and solve (22) for y(x) and u(x). Next, we compare the computed value of $y(5)$ with the boundary condition $y(5) = 2$ (which we have not yet used). If the computed value is too high, then we return to (22), reduce the value of u_0, and solve again. Comparing the new computed value of $y(5)$ with the prescribed value $y(5) = 2$, we again revise our value of u_0. If these revisions are done in a rational way, one can imagine obtaining a convergent scheme. Such a scheme is called a *shooting method* because of the obvious analogy with the shooting of a projectile, with the intention of having the projectile strike the ground at some distant prescribed point.

Thus, we can see the increase in difficulty as we move away from linear initial-value problems. For a linear boundary-value problem of order n we need to solve not one problem but $n+1$ of them. For a nonlinear boundary-value problem we need to solve an infinite sequence of them, in principle; in practice, we need to carry out only enough iterations to produce the desired accuracy.

Closure. In Section 6.4.1 we extend the Euler and fourth-order Runge–Kutta solution methods to cover systems of equations and higher-order equations. In that discussion it is more convenient to use n-dimensional vector notation because of its compactness, but that notation is not be introduced until Chapters 9 and 10. Nonetheless, let us indicate the result, if only for the Euler method, for future reference. The idea is that we can express the system

$$\begin{aligned} y_1'(x) &= f_1\left(x, y_1(x), \dots, y_n(x)\right); \quad & y_1(a) &= y_{10}, \\ &\vdots & &\vdots \\ y_n'(x) &= f_n\left(x, y_1(x), \dots, y_n(x)\right); \quad & y_n(a) &= y_{n0}, \end{aligned} \tag{23}$$

in the vector form

$$\mathbf{y}'(x) = \mathbf{f}\left(x, \mathbf{y}(x)\right); \qquad \mathbf{y}(a) = \mathbf{y_0}, \tag{24}$$

where the boldface letters denote "n-dimensional column vectors:"

$$\mathbf{y}(x) = \begin{bmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{bmatrix}, \quad \mathbf{y}'(x) = \begin{bmatrix} y_1'(x) \\ \vdots \\ y_n'(x) \end{bmatrix}, \quad \mathbf{f}(x, \mathbf{y}(x)) = \begin{bmatrix} f_1(x, \mathbf{y}(x)) \\ \vdots \\ f_n(x, \mathbf{y}(x)) \end{bmatrix}, \tag{25}$$

and where $f_j(x, \mathbf{y}(x))$ is simply a shorthand notation for $f_j(x, y_1(x), \dots, y_n(x))$. Then the Euler algorithm corresponding to (24) is

$$\mathbf{y_{n+1}} = \mathbf{y_n} + \mathbf{f}\left(x_n, \mathbf{y_n}\right) h. \tag{26}$$

In Section 6.4.2 we turn to boundary-value problems, but only linear ones. Given an nth-order linear boundary-value problem $L[y] = f(x)$ on an interval $[a, b]$ plus n boundary conditions the idea is to solve the problems

$$\begin{aligned} &L\left[Y_1\right] = 0; \quad Y_1(a) = 1, \ Y_1'(a) = \cdots = Y_1^{(n-1)}(a) = 0, \\ &\quad \vdots \qquad\qquad\qquad\qquad \vdots \\ &L\left[Y_n\right] = 0; \quad Y_n(a) = Y_n'(a) = \cdots = Y_n^{(n-2)}(a) = 0, \ Y_n^{(n-1)}(a) = 1, \\ &L\left[Y_p\right] = f; \quad Y_p(a) = \cdots = Y_p^{(n-1)}(a) = 0 \end{aligned} \tag{27}$$

for $Y_1(x), \ldots, Y_n(x), Y_p(x)$ and to form the general solution as

$$y(x) = C_1Y_1(x) + \cdots + C_nY_n(x) + Y_p(x). \tag{28}$$

Finally, application of the original boundary conditions to (28) yields n linear algebraic equations for $C_1, \ldots, C_n$, which equations will have a unique solution, no solution, or an infinity of solutions.

Computer software. No new software is needed for the methods described in this section. For instance, we can use the *Maple* command dsolve, with the numeric option, to solve the problem

$$\begin{aligned} u' &= x + v; \quad u(0) = 0 \\ v' &= -5uv; \quad v(0) = 1 \end{aligned}$$

and to print the results at $x = 1, 2$, and 3. First, enter

```
with(DEtools):
```

and return. Then enter

```
dsolve({diff(u(x), x) = x + v(x), diff(v(x), x) = -5 * u(x) * v(x),
u(0) = 0, v(0) = 1}, {u(x), v(x)}, type = numeric,
value = array ([1, 2, 3]));
```

and return. The printed result is

$$\left[\begin{array}{c} [x, u(x), v(x)] \\ \left[\begin{array}{lll} 1. & 1.032499017614234 & .07285274036469075 \\ 2. & 2.544584704578166 & .00001413488345836790 \\ 3. & 5.044585755162072 & -.3131443346304622 \times 10^{-9} \end{array}\right] \end{array}\right]$$

The only differences between the command above and the one given at the end of Section 6.3 is that here we have entered two differential equations, two initial conditions, and two dependent variables, and we have omitted the abserr option.

Observe that to solve a differential equation, or system of differential equations, numerically, we must first express the equations as a system of first-order equations, as illustrated in Example 4. However, to use the *Maple* dsolve command we can leave the original higher-order equations intact.

EXERCISES 6.4

1. In Example 2 we gave $k_1, l_1, k_2, l_2, k_3, l_3, k_4, l_4$ for $n = 1$, and the resulting values of u_2 and v_2, but did not show the calculations. Provide those calculations, as we did for the step $n = 0$.

2. As we did in Example 1, work out y_1, z_1, by hand. Use three methods: Euler, second-order Runge–Kutta, and fourth-order Runge–Kutta, and take $h = 0.2$. These problems are rigged so as to have simple closed-form solutions, some of which are given in brackets. Compare your results with the exact solution.

(a) $y' = z; \quad y(0) = 1$
$z' = -y; \quad z(0) = 0$

(b) $y' = 4z; \quad y(2) = 5$
$z' = -y; \quad z(2) = 0$

(c) $y' = -z^3/y; \quad y(0) = 1 \quad [y(x) = e^{-x}]$
$z' = -y; \quad z(0) = 0 \quad [z(x) = e^{-x}]$

(d) $y' = 2xz^2/y; \quad y(1) = 1 \quad [y(x) = x^2]$
$z' = y/z^2; \quad z(1) = 1 \quad [z(x) = x]$

(e) $y' = (x+y)z - 1; \quad y(1) = 1 \quad [y(x) = x]$
$z' = -yz^3; \quad z(1) = 1 \quad [z(x) = 1/x]$

3. (a)–(e) First, read Exercise 2. Use computer software to solve the initial-value problem given in the corresponding part of Exercise 2, for $y(x)$ and $z(x)$ at $x = 3, 5$, and 10, and compare those results with the exact solution at those points.

4. (a) Just as (2) and (6) give the Euler and fourth-order Runge-Kutta algorithms for the second-order system (1), write down the analogous Euler, second-order Runge–Kutta, and fourth-order Runge–Kutta algorithms for the third-order system

$$\begin{aligned} x'(t) &= f(t,x,y,z), \quad x(a) = x_0 \\ y'(t) &= g(t,x,y,z), \quad y(a) = y_0 \qquad (4.1) \\ z'(t) &= f(t,x,y,z), \quad z(a) = z_0. \end{aligned}$$

Use the Euler and second-order Runge–Kutta algorithms to work out x_1, y_1, z_1 and x_2, y_2, z_2, by hand, for the case where f, g, h are $y - 1, z, t + x + 3(z - y + 1)$, respectively, with the initial conditions $x(0) = -3$, $y(0) = 0$, $z(0) = 2$ using $h = 0.3$.
(b) Same as (a), but with $x(0) = y(0) = z(0) = 0$.
(c) Same as (a), but with $x(0) = 1, y(0) = 0, z(0) = 0$
(d) Same as (a), but with $x(0) = y(0) = 0, z(0) = 10$.

5. We re-expressed (8) and (10) as the equivalent systems of first-order initial-value problems (9) and (11), respectively. Do the same for the problem given. You need go no further.

(a) $mx'' + cx' + kx = f(t); \quad x(0) = x_0, \quad x'(0) = x'_0$
(b) $Li'' + Ri' + (1/C)i = E'(t); \quad i(0) = i_0, \quad i'(0) = i'_0$
(c) $y'' - xyy' = \sin x; \quad y(1) = 5, \quad y'(1) = -1$
(d) $y'' + y' - 4y = 3x; \quad y(0) = 2, \quad y'(0) = 1$
(e) $y''' - 2\sin y' = 3x; \quad y(-2) = 7, \quad y'(-2) = 4$
$y''(-2) = 0$
(f) $y''' + xy = \cos 2x; \quad y(1) = 3, \quad y'(1) = 2, \quad y''(1) = 0$
(g) $x'' + 2x - 3y = 10\cos 3t; \quad x(0) = 2, \quad x'(0) = -1$
$y'' - x + 5y = 0; \quad y(0) = 4, y'(0) = 3$
(h) $y''' + xy'z = f(x); \quad y(3) = 2, \quad y'(3) = -1$
$y''(3) = 6$
$z'' + y - z = g(x); \quad z(3) = 0, z'(3) = 8$
(i) $x'' - 3xz = \sin t; \quad x(1) = x'(1) = 0, \quad x''(1) = 3$
$y' + 2x + y - 5z = 0; \quad y(1) = 6$
$z' - 4xy = e^{2t}; \quad z(1) = 2$
(j) $y''' = yz; \quad y(0) = 1, \quad y'(0) = y''(0) = 0$
$z'' = -xy + z; \quad z(0) = 5, z'(0) = 1$

6. Use computer software to solve the given system numerically, and print out the solution for $y(x)$ and $z(x)$ at $x = 1, 2$.

(a) $y'' - 2xz' = 5x; \quad y(0) = 2, \quad y'(0) = -1$
$z' + yz^3 = -3; \quad z(0) = 1$

(b) $y' + 3xz = x^2; \quad y(0) = 1$
$z''' - y^2z' + z = 0; \quad z(0) = z'(0) = 2, \quad z''(0) = -1$

(c) $y' = z' + x; \quad y(2) = 1$
$z'' + y^2z^2 = 3; \quad z(2) = -1, \quad z'(2) = 0$

(d) $y' - z'' = x; \quad y(1) = -1$
$z''' - y^2z = 3; \quad z(1) = 1, \quad z'(1) = z''(1) = 0$

7. Complete the solution of Example 5 by using computer software to solve (18) for $Y_3(x)$ and (19) for $Y_p(x)$, at $x = 2, 4, 6$, and then using (17) to determine $y(x)$ at those points.

8. Use the method explained in Example 5 to reduce the given linear boundary-value problem to a system of linear initial-value problems. Then complete the solution and solve for the specified quantity, either using computer software or by programming any of the numerical solution methods that we have studied. Obtain accuracy to five significant figures, and indicate why you believe that you have achieved that accuracy.

If you believe that there is no solution or that it exists but is nonunique, then give your reasoning. HINT: You can specify homogeneous initial conditions for the Y_p problem, as we did in Example 5, but be aware that you do not *have* to use homogeneous conditions, and that you may be able to reduce your labor by a more optimal choice of those conditions.

(a) $y'' - 2xy' + y = 3\sin x;\quad y(0) = 1,\quad y(2) = 3.$
Determine $y(1)$.

(b) $y'' + (\cos x)y = 0;\quad y(0) = 1,\quad y(10) = 2.$
Determine $y(2)$.

(c) $y'' - [\ln(x+1)]\,y' - y = 2\sin 3x + 1;$
$y(0) = 3,\quad y(2) = -1.$
Determine $y(x)$ at $x = 0.5, 1.0, 1.5$.

(d) $y'' + y' - xy = x^3;\quad y(0) = 1,\quad y(5) = 2.$
Determine $y(x)$ at $x = 1, 2, 3, 4$.

(e) $y''' + xy = 2x^3;\quad y(1) = y'(1) = 0,\quad y''(2) = -3.$
Determine $y(2)$.

(f) $y''' + xy' + y = x;\ y(1) = 2,\ y'(1) = 0.4,\ y'(5) = 3.$
Determine $y(x)$ at $x = 4, 5$.

6.5 Stability and Difference Equations

6.5.1. Introduction. In progressing from the simple Euler method to the more sophisticated higher-order methods our aim was improvement in accuracy. However, there are cases where the results obtained not only fail to be sufficiently accurate but are grossly incorrect, as illustrated in the two examples to follow. The second one introduces the idea of stability, and in Section 6.5.2 we concentrate on that topic.

EXAMPLE 1. The initial-value problem

$$y' - 2y = -6e^{-4x};\quad y(0) = 1 \tag{1}$$

has the exact solution $y(x) = \exp(-4x)$. If we solve it by the fourth-order Runge–Kutta method for the step sizes $h = 0.1, 0.05$, and 0.01, we obtain in Table 1 the results shown, at

Table 1. Runge–Kutta solution of (1).

x	$h = 0.1$	$h = 0.05$	$h = 0.01$	Exact
1	0.179006 E−1	0.182893 E−1	0.183153 E−1	0.183156 E−1
4	−0.167842 E+0	−0.106538 E−1	−0.146405 E−3	0.112535 E−6
8	−0.500286 E+3	−0.317586 E+2	−0.436763 E+0	0.126642 E−13
12	−0.149120 E+7	−0.946704 E+5	−0.130197 E+4	0.142516 E−20

the representative points $x = 1, 4, 8$, and 12. Since the Runge–Kutta method is convergent, the results should converge to the exact solution at any given x as h tends to zero, but that convergence is hard to see in the tabulated results except for $x = 1$. In fact, it is doubtful that we could ever come close to the exact values at $x = 8$ or 12 since we might need to make h so small that roundoff errors might come to dominate before the accumulated truncation error is sufficiently reduced.

More central to the purpose of this example is to see that with h fixed the results diverge dramatically from the exact solution as x increases so as to become grossly incorrect. We cannot blame this strange and unexpected result on complications due to nonlinearity because (1) is linear.

To understand the source of the difficulty, note that the general solution of the differential equation is $y(x) = \exp(-4x) + C\exp(2x)$, where C is an arbitrary constant. The initial condition implies that $C = 0$, leaving the particular solution $y(x) = \exp(-4x)$. In Fig. 1 we show several solution curves for values of C close to and equal to zero, and we can see the rapid divergence of neighboring curves from the solution $y(x) = \exp(-4x)$. Thus, the explanation of the difficulties found in the tabulated numerical results is that even a very small numerical error shifts us from the exact solution curve to a neighboring curve, which then diverges from the true solution. ∎

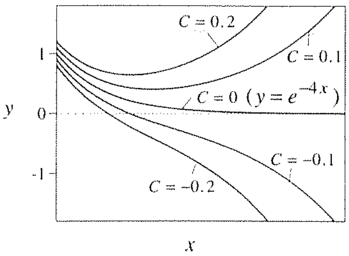


Figure 1. Solution curves for the equation $y' - 2y = -6e^{-4x}$.

EXAMPLE 2. In Example 3 of Section 6.3 we solved the equation $y' = -y$, with initial condition $y(0) = 1$, by several methods – from the simple Euler method to the more accurate and sophisticated fourth-order Runge–Kutta method, and we gave the results in Table 2. Since the midpoint rule and the second-order Runge–Kutta methods are both of second order we expected their accuracy to be comparable. Indeed they were initially, but the midpoint rule eventually developed an error that oscillated in sign from step to step and grew in magnitude (see Table 2 in Section 6.3). Let us solve the similar problem

$$y' = -2y; \quad y(0) = 1 \tag{2}$$

by the midpoint rule, with $h = 0.05$. Since the midpoint rule is not self-starting, we use ten Euler steps from $x = 0$ to $x = 0.05$ before switching over to the midpoint rule. We have plotted the results in Fig. 2, along with the exact solution, $y(x) = \exp(-2x)$. Once again, we see that the midpoint rule results follow the exact solution initially, but they develop an error that oscillates in sign and grows such that the results are soon hopelessly incorrect.

This numerical difficulty is different from the one found above in Example 1, for rather than being due to an extreme sensitivity to initial conditions, it is associated with machine roundoff error and is an example of numerical instability. ∎

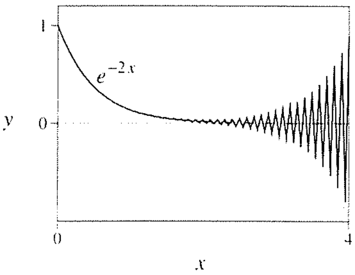


Figure 2. Illustration of numerical instability associated with the midpoint rule, for the initial-value problem (2).

6.5.2. Stability. Let us analyze the phenomenon of numerical instability that we encountered in Example 2. Recall that we denote the exact solution of a given initial-value problem as $y(x_n)$ and the numerical solution as y_n. Actually, the latter is not quite the same as the computer printout because of the inevitable presence of machine roundoff errors. Thus, let us distinguish further between the numerical solution y_n that would be generated on a perfect computer, and the solution y_n^* that is generated on a real machine and which includes the effects of numerical roundoff – that is, the truncation of numbers after a certain number of significant figures.

It is useful to decompose the total error, at any nth step, as

$$\boxed{\begin{aligned}\text{Total error} &= y(x_n) - y_n^* = [y(x_n) - y_n] + [y_n - y_n^*] \\ &= [\text{accum. truncation error}] + [\text{accum. roundoff error}].\end{aligned}} \tag{3}$$

We ask two things of a method: first, that the accumulated truncation error tend to zero at any fixed x as the step size h tends to zero and, second, that the accumulated roundoff error remain small compared to the exact solution. The first is the issue of **convergence**, discussed earlier in this chapter, and the second is the issue of **stability**, our present concern.

We have already noted that the midpoint rule can produce the strange behavior shown in Fig. 2, so let us study the application of that method to the standard "test problem,"

$$y' = Ay; \quad y(0) = 1, \tag{4}$$

where it is useful to include the constant A as a parameter. The midpoint rule generates y_n according to the algorithm

$$\begin{aligned} y_{n+1} &= y_{n-1} + 2hf(x_n, y_n) \\ &= y_{n-1} + 2Ay_n; \qquad y_0 = 1 \end{aligned} \tag{5}$$

for $n = 0, 1, 2, \dots$.

To determine whether a solution algorithm, in this case (5), is stable, it is customary to "inject" a roundoff error at any step in the solution, say at $n = 0$, and to see how much the perturbed solution differs from the exact solution as n increases, assuming that no further roundoff errors occur. Thus, in place of (5), consider the perturbed problem

$$y^*_{n+1} = y^*_{n-1} + 2Ahy^*_n; \quad y^*_0 = 1 - \epsilon, \tag{6}$$

say, where ϵ is the (positive or negative) roundoff error in the initial condition. Defining the error $e_n \equiv y_n - y^*_n$ and subtracting (6) from (5), gives

$$e_{n+1} = e_{n-1} + 2Ahe_n, \tag{7}$$

with the initial condition $e_0 = \epsilon$, as governing the evolution of e_n. We call (7) a **difference equation**. Just as certain differential equations can be solved by seeking solutions in the form $y(x) = e^{\lambda x}$, the appropriate form for the difference equation (7) is

$$e_n = \rho^n, \tag{8}$$

where ρ is to be determined. Putting this expression into (7) gives

$$\rho^{n+1} - 2Ah\rho^n - \rho^{n-1} = 0 \tag{9}$$

or

$$\left(\rho^2 - 2Ah\rho - 1\right)\frac{1}{\rho^n} = 0. \tag{10}$$

Since $1/\rho^n$ is not zero, it follows from (10) that we must have $\rho^2 - 2Ah\rho - 1 = 0$, so we have the two roots

$$\rho = Ah + \sqrt{1 + A^2h^2} \quad \text{and} \quad \rho = Ah - \sqrt{1 + A^2h^2}. \tag{11}$$

By considerations analogous to those for differential equations, we have

$$e_n = C_1 \left(Ah + \sqrt{1 + A^2h^2} \right)^n + C_2 \left(Ah - \sqrt{1 + A^2h^2} \right)^n \tag{12}$$

as the general solution of (7).

If we let $h \to 0$, then

$$\left(Ah + \sqrt{1 + A^2h^2} \right)^n \sim (Ah + 1)^n = e^{n \ln (1+Ah)} \sim e^{nAh} = e^{Ax_n}, \tag{13}$$

where we have used the identity $a = e^{\ln a}$, the Taylor expansion $\ln (1 + x) = x - x^2/2 + \cdots \sim x$, and the fact that $x_n = nh$. Similarly,

$$\begin{aligned} \left(Ah - \sqrt{1 + A^2h^2} \right)^n &\sim (Ah - 1)^n \\ &= (-1)^n (1 - Ah)^n = (-1)^n e^{n \ln (1-Ah)} \\ &\sim (-1)^n e^{-nAh} = (-1)^n e^{-Ax_n}, \end{aligned} \tag{14}$$

so (12) becomes

$$e_n = C_1 e^{Ax_n} + C_2 (-1)^n e^{-Ax_n} \tag{15}$$

as $h \to 0$. Since there are two arbitrary constants, C_1 and C_2, two initial conditions are appropriate, whereas we have attached only the single condition $e_0 = \epsilon$ in (7). With no great loss of generality let us specify as a second initial condition $e_1 = 0$. Imposing these conditions on (15), we have

$$\begin{aligned} e_0 &= \epsilon = C_1 + C_2, \\ e_1 &= 0 = C_1 e^{4x_1} - C_2 e^{-4x_1}. \end{aligned}$$

Finally, solving for C_1 and C_2 and inserting these values into (15), gives

$$e_n = \frac{\epsilon}{2 \cosh Ax_1} \left[e^{A(x_n - x_1)} + (-1)^n e^{-A(x_n - x_1)} \right]. \tag{16}$$

To infer from (16) whether the method is stable or not, we consider the cases $A > 0$ and $A < 0$ separately. If $A > 0$, then the second term in (16) decays to zero, and even though the first term grows exponentially, it remains small compared to the exact solution $y(x_n) = \exp (Ax_n)$ as n increases because ϵ is very small (for example, on the order of 10^{-10}). We conclude, formally, that if $A > 0$ then the midpoint rule is stable.

On the other hand, if $A < 0$, then the second term starts out quite small, due to the ϵ factor, but grows exponentially with x_n and oscillates due to the $(-1)^n$, whereas the exact solution is $\exp (-x)$. This is precisely the sort of behavior that was observed in Example 2 (where A was -2), and we conclude that if $A < 0$, then the midpoint rule is unstable.

Since the stability of the midpoint rule depends upon the sign of A in the test equation $y' = Ay$ (stability for $A > 0$ and instability for $A < 0$), we say that the midpoint rule is only **weakly stable**. If, instead, a method is stable independent of the sign of A, then we classify it as **strongly stable**.

Having found that the midpoint rule when applied to the equation $y' = Ay$ is stable for $A > 0$ and unstable for $A < 0$, what about the stability of the midpoint rule when it is applied to an equation $y' = f(x, y)$ that is more complicated? Observing that A is the partial derivative of Ay (i.e., the right-hand side of $y' = Ay$) with respect to y, we expect, as a rule of thumb, the midpoint rule to be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the x, y domain of interest. For instance, if $y' = e^{xy}$ on $x > 0$, then we can expect the midpoint rule to be stable because $\partial(e^{xy})/\partial y = xe^{xy} > 0$ on $x > 0$, but if $y' = e^{-xy}$ on $x > 0$, then we can expect the midpoint rule to be unstable on $x > 0$ because $\partial(e^{-xy})/\partial y = -xe^{-xy} < 0$ on $x > 0$.

Besides arriving at the above-stated conclusions as to the stability of the midpoint rule for the test equation $y' = Ay$, we can now understand the origin of the instability, for notice that the difference equations $y_{n+1} - 2Ahy_n - y_{n-1} = 0$ and $e_{n+1} - 2Ahe_n - e_{n-1} = 0$, governing y_n and e_n, are identical. Thus, analogous to (15) we must have

$$y_n = B_1 e^{Ax_n} + B_2(-1)^n e^{-Ax_n} \tag{17}$$

for arbitrary constants B_1, B_2, as h tends to zero. The first of these terms coincides with the exact solution of the original equation $y' = Ay$, and the second term (which gives rise to the instability if $A < 0$) is an extraneous solution that enters because we have replaced the original first-order differential equation by a second-order difference equation (second-order because the difference between the subscripts $n+1$ and $n-1$ is 2). Single-step methods (e.g., Euler and Runge–Kutta) are strongly stable (i.e., independent of the sign of A) because the resulting difference equation is only first order so there are no extraneous solutions. Thus, we can finally see why, in Example 3 of Section 6.3, the midpoint rule proved unstable but the other methods were stable.

Understand that these stability claims are based upon analyses in which we let h tend to zero, whereas in practice h is, of course, finite. To illustrate what can happen as h is varied, let us solve

$$y' = -1000(y - x^3) + 3x^2; \quad y(0) = 0 \tag{18}$$

by Euler's method. The exact solution is simply $y(x) = x^3$ so that at $x = 1$, for instance, we have $y(1) = 1$. By comparison, the values computed by Euler's method are as given in Table 2.

Even from this limited data we can see that we do have the stability claimed above for the single-step Euler method, but only when h is made sufficiently small. To understand this behavior, consider the relevant test equation $y' = -1000y$, namely, $y' = Ay$, where $A = \partial[-1000(y - x^3) + 3x^2]/\partial y = -1000$. Then Euler's method for that test equation is $y_{n+1} = y_n - 1000hy_n$. Similarly, $y^*_{n+1} = y^*_n - 1000hy^*_n$. Subtracting these two equations, we find that the roundoff error $e_n = y_n - y^*_n$ satisfies the simple difference equation

$$e_{n+1} = (1 - 1000h)e_n. \tag{19}$$

Table 2. Finite-h stability.

h	Computed $y(1)$
0.2500	2.3737566×10^{5}
0.1000	8.7725049×10^{14}
0.0100	Exponential overflow
0.0010	0.99999726
0.0001	0.99999970

Letting $n = 0, 1, 2, \dots$ in (19) reveals that the solution of (19) is

$$e_n = (1 - 1000h)^n e_0, \tag{20}$$

where e_0 is the initial roundoff error. If we take the limit as $h \to 0$, then

$$e_n = (1 - 1000h)^n e_0 = e_0 e^{n \ln(1-1000h)} \sim e_0 e^{-1000nh} = e_0 e^{-1000x_n}, \tag{21}$$

which is small compared to the exact solution $y_n = e^{-1000x_n}$ because of the e_0 factor, so the method is stable. This result is in agreement with the numerical results given in Table 2: as $h \to 0$ the scheme is stable. However, in a real calculation h is, of course, finite and it appears, from the tabulation that there is some critical value, say h_{cr}, such that the guaranteed stability is realized only if $h < h_{cr}$. To see this, let us retain (20) rather than let $h \to 0$. It is seen from (20) that if $|1 - 1000h| < 1$, then $e_n \to 0$ as $n \to \infty$, and if $|1 - 1000h| > 1$, then $e_n \to \infty$ as $n \to \infty$. Thus, for stability we need $|1 - 1000h| < 1$ or $-1 < 1 - 1000h < 1$. The right-hand inequality imposes no restriction on h because it is true for all h's (provided that h is positive, as is normally the case), and the left-hand inequality is true only for $h < 0.002$. Hence $h_{cr} = 0.002$ in this example, and this result is consistent with the tabulated results, which show instability for the h's greater than that value, and stability for the h's smaller than that value. Thus, when we say that the Euler method is strongly stable, what we should really say is that it is strongly stable for *sufficiently small* h. Likewise for the Runge–Kutta and other single-step methods.

6.5.3. Difference equations. (Optional) Difference equations are important in their own right, and the purpose of this Section 6.5.3 is not only to clarify some of the steps in Section 6.5.2, but also to take this opportunity to present some basics regarding the theory and solution of such equations.

To begin, we define a **difference equation of order** N as a relation involving $y_n, y_{n+1}, \dots, y_{n+N}$. As we have seen, one way in which difference equations arise is in the numerical solution of differential equations. For instance, if we discretize the differential equation $y' = -y$ and solve by Euler's method or the midpoint rule, then in place of the differential equation we have the first- and second-order difference equations $y_{n+1} = y_n - hy_n = (1 - h)y_n$ and $y_{n+1} = y_{n-1} - 2hy_n$, or

$$y_{n+1} - (1 - h)y_n = 0 \tag{22}$$

and

$$y_{n+1} + 2hy_n - y_{n-1} = 0, \tag{23}$$

respectively. In case it is not clear that (23) is of second order, we could let $n-1 = m$ and obtain $y_{m+2} + 2hy_{m+1} - y_m = 0$ instead, which equation is more clearly of second order. That is, the order is always the difference between the largest and smallest subscripted indices.

Analogous to differential equation terminology, we say that (22) and (23) are **linear** because they are of the form

$$a_0(n)y_{n+N} + a_1(n)y_{n+N-1} + \cdots + a_N(n)y_n = f(n), \tag{24}$$

homogeneous because $f(n)$ is zero in each case, and of **constant-coefficient** type because their a_j's are constants rather than functions of n. By a solution of (24) is meant any sequence y_n that reduces (24) to a numerical identity for each n under consideration, such as $n = 0, 1, 2, \ldots$.

The theory of difference equations is analogous to that of differential equations. For instance, just as one seeks solutions to a linear homogeneous differential equation with constant coefficients in the form $y(x) = e^{\lambda x}$, one seeks solutions to a linear homogeneous difference equation with constant coefficients in the form

$$y_n = \rho^n \tag{25}$$

as we did in Section 6.5.2. [In case these forms don't seem analogous, observe that $e^{\lambda x} = (e^{\lambda})^x$ is a constant to the power x, just as ρ^n is a constant to the power n.] Putting (25) into such an Nth-order difference equation gives an Nth-degree polynomial equation on ρ, the characteristic equation corresponding to the given difference equation, and if the N roots $(\rho_1, \ldots, \rho_N)$ are distinct, then

$$y_n = C_1\rho_1^n + \cdots + C_N\rho_2^N, \tag{26}$$

where the C_j's are arbitrary constants, can be shown to be a **general solution** of the difference equation in the sense that every solution is of the form (26) for some specific choice of the C_j's. For an Nth-order linear differential equation, N initial conditions (y and its first $N-1$ derivatives at the initial point) are appropriate for narrowing a general solution down to a particular solution. Likewise for a linear difference equation N initial conditions are appropriate – namely, the first N values $y_0, y_1, \ldots, y_{N-1}$.

EXAMPLE 3. Solve the difference equation

$$y_{n+1} - 4y_n = 0. \tag{27}$$

Since (27) is linear, homogeneous and of constant-coefficient type, seek solutions in the form (25). Putting that form into (27) gives

$$\rho^{n+1} - 4\rho^n = (\rho - 4)\rho^n = 0 \tag{28}$$

so that if $\rho \neq 0$ then $\rho - 4 = 0$, $\rho = 4$, and the general solution of (27) is

$$y_n = C(4)^n. \tag{29}$$

For example, if an initial condition $y_0 = 3$ is specified, then $y_0 = 3 = C(4)^0 = C$ gives $C = 3$ and hence the particular solution $y_n = 3(4)^n$.

Actually, (29) is simple enough to solve more directly since (for $n = 0, 1, \ldots$) it gives $y_1 = 4y_0$, $y_2 = 4y_1 = 4^2 y_0$, $y_3 = 4y_2 = 4^3 y_0$, and so on, so one can see that $y_n = y_0(4)^n$ or, if y_0 is not specified, $y_n = C(4)^n$, which is the same as (29). ▮

EXAMPLE 4. Solve the difference equation

$$y_{n+2} - y_{n+1} - 6y_n = 0. \tag{30}$$

Seeking solutions in the form (25) gives the characteristic equation $\rho^2 - \rho - 6 = 0$ with roots -2 and 3 so the general solution of (30) is

$$y_n = C_1(-2)^n + C_2(3)^n. \tag{31}$$

If initial conditions are prescribed, say $y_0 = 4$ and $y_1 = -13$, then

$$\begin{aligned} y_0 &= 4 = C_1 + C_2, \\ y_1 &= -13 = -2C_1 + 3C_2 \end{aligned}$$

give $C_1 = 5$ and $C_2 = -1$. ▮

If the characteristic polynomial has a pair of complex conjugate roots, say $\rho_1 = \alpha + i\beta$ and $\rho_2 = \alpha - i\beta$, then the solution can still be expressed in real form, for if we express ρ_1 and ρ_2 in polar form (as is explained in Section 22.2 on complex numbers) as

$$\rho_1 = r\,e^{i\theta} \quad \text{and} \quad \rho_2 = r\,e^{-i\theta}, \tag{32}$$

where $r = \sqrt{\alpha^2 + \beta^2}$ and $\theta = \tan^{-1}(\beta/\alpha)$, then

$$\begin{aligned} C_1\rho_1^n + C_2\rho_2^n &= C_1 r^n e^{in\theta} + C_2 r^n e^{-in\theta} \\ &= r^n \left(C_1 e^{in\theta} + C_2 e^{-in\theta} \right) \\ &= r^n \left[C_1 (\cos n\theta + i \sin n\theta) + C_2 (\cos n\theta - i \sin n\theta) \right] \\ &= r^n \left(C_3 \cos n\theta + C_4 \sin n\theta \right), \end{aligned} \tag{33}$$

where $C_3 = C_1 + C_2$ and $C_4 = i(C_1 - C_2)$ are arbitrary constants.

EXAMPLE 5. Solve the difference equation

$$y_{n+2} + 4y_n = 0. \tag{34}$$

The characteristic roots are $\pm 2i$ so (32) becomes $\rho_1 = 2e^{i\pi/2}$ and $\rho_2 = 2e^{-i\pi/2}$, and (33) gives the general solution

$$y_n = 2^n \left(A \cos \frac{n\pi}{2} + B \sin \frac{n\pi}{2} \right), \tag{35}$$

where A, B are arbitrary constants. ▮

As we have seen, one way in which difference equations arise is in the numerical solution of differential equations, wherein the continuous process (described by the differential equation) is approximated by a discrete one (described by the difference equation). However, they also arise directly in modeling discrete processes. To illustrate, let p_n be the principal in a savings account at the end of the nth year. We say that the process is discrete in that p is a function of the integer variable n rather than a continuous time variable t. If the account earns an annual interest of I percent, then the growth of principal from year to year is governed by the difference equation

$$p_{n+1} = \left(1 + \frac{I}{100} \right) p_n, \tag{36}$$

which is of the same form as (22).

In fact, discrete processes governed by nonlinear difference equations are part of the modern theory of *dynamical systems*, in which theory the phenomenon of chaos plays a prominent role. Let us close with a brief mention of one such discrete process that is familiar to those who study dynamical systems and chaos. Let x_n, y_n be a point in a Cartesian x, y plane, and let its polar coordinates be r and θ_n. Consider a simple process, or **mapping**, which sends that point into a point x_{n+1}, y_{n+1} at the same radius but at an incremented angle $\theta_n + \alpha$. Then, recalling the identity $\cos(A+B) = \cos A \cos B - \sin A \sin B$, we can express $x_{n+1} = r\cos(\theta_n + \alpha) = r\cos\theta_n \cos\alpha - r\sin\theta_n \sin\alpha = x_n \cos\alpha - y_n \sin\alpha$ and, recalling the identity $\sin(A+B) = \sin A \cos B + \sin B \cos A$, we can express $y_{n+1} = r\sin(\theta_n + \alpha) = r\sin\theta_n \cos\alpha + r\cos\theta_n \sin\alpha = y_n \cos\alpha + x_n \sin\alpha$. Thus, the process is described by the system of linear difference equations

$$\begin{aligned} x_{n+1} &= (\cos\alpha)x_n - (\sin\alpha)y_n, \\ y_{n+1} &= (\sin\alpha)x_n + (\cos\alpha)y_n. \end{aligned} \tag{37}$$

Surely, if one plots such a sequence of points it will fall on the circle of radius r centered at the origin. Suppose that we now modify the process by including two quadratic x_n^2 terms, so that we have the *non*linear system

$$\begin{aligned} x_{n+1} &= (\cos\alpha)x_n - (\sin\alpha)\left(y_n - x_n^2\right), \\ y_{n+1} &= (\sin\alpha)x_n + (\cos\alpha)\left(y_n - x_n^2\right). \end{aligned} \tag{38}$$

This system, studied initially by *M. Henòn*, turns out to be remarkably complex and interesting by virtue of the nonlinearity. For a discussion of the main results, we highly recommend the little book *Mathematics and the Unexpected*, by Ivar Ekeland (Chicago: University of Chicago Press, 1988).

Closure. This section is primarily about the concept of stability in the numerical solution of differential equations. A scheme is stable if the roundoff error remains small compared to the exact solution. Normally, one establishes the stability or instability of a method with respect to the simple test equation $y' = Ay$. Assuming that roundoff enters in the initial condition and that the computer is perfect thereafter, one can derive a difference equation governing the roundoff error e_n, and solve it analytically to see if e_n remains small. Doing so, we show that the midpoint rule is only weakly stable: stable if $A > 0$ and unstable if $A < 0$. As a rule of thumb, we suggest that for a given differential equation $y' = f(x, y)$ we can expect the midpoint rule to be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the x, y region of interest.

To explain the source of the instability in the midpoint rule, we observe that the exact solution (17) of the midpoint rule difference equation corresponding to the test equation $y' = Ay$ contains two terms, one that corresponds to the exact solution of $y' = Ay$ and the other extraneous. The latter enters because the midpoint rule difference equation is of second order, whereas the differential equation is only of first order, and it is that extraneous term that leads to the instability. Single-step methods such as the Euler and Runge–Kutta methods, however, are strongly stable, provided that h is sufficiently small.

Observe that the only multi-step method that we examine is the midpoint rule; we neither show nor claim that all multi-step methods exhibit such instability. For instance, it is left for the exercises to show that the multi-step fourth-order Adams–Moulton method is strongly stable (for sufficiently small h). Thus, the idea is that the extraneous terms in the solution, that arise because the difference equation is of a higher order than the differential equation, can, but need not, cause trouble.

We close the section with a brief study of difference equations, independent of any connection with differential equations and stability since they are important in their own right in the modeling of discrete systems. We stress how analogous are the theories governing differential and difference equations that are linear, homogeneous, and with constant coefficients.

Computer software. Just as many differential equations can be solved analytically by computer-algebra systems, so can many difference equations. Using *Maple*, for instance, the relevant command is **rsolve**. For instance, to solve the difference equation $y_{n+2} - y_{n+1} - 6y_n = 0$ (from Example 4), enter

$$\text{rsolve}(y(n+2) - y(n+1) - 6 * y(n) = 0,\ y(n));$$

and return. The result is

$$\left(\frac{3}{5}y(0) - \frac{1}{5}y(1)\right)(-2)^n - \left(-\frac{2}{5}y(0) - \frac{1}{5}y(1)\right)(3)^n$$

the correct solution for any initial values $y(0)$ and $y(1)$. Of course, we could re-express the latter as

$$C_1(-2)^n + C_2(3)^n$$

if we wish. If we had initial conditions $y_0 = 4$ and $y_1 = -7$, say, then we would have entered

$$\text{rsolve}(\{y(n+2) - y(n+1) - 6 * y(n) = 0,\ y(0) = 4,\ y(1) = -7\},$$
$$y(n));$$

and would have obtained

$$\frac{19}{5}(-2)^n + \frac{1}{5}(3)^n$$

as the desired particular solution.

EXERCISES 6.5

1. If the given initial-value problem were to be solved by the fourth-order Runge–Kutta method (and we are not asking you to do that), do you think accurate results could be obtained? Explain. The x domain is $0 \leq x < \infty$.

(a) $y' = 2y - 8x + 4;\quad y(0) = 0$
(b) $y' = y - 2e^{-x};\quad y(0) = 1$
(c) $y' = y + 5e^{-4x};\quad y(0) = -1$
(d) $y' = 1 + 3(y - x);\quad y(0) = 0$

2. It is natural to wonder how well we would fare trying to solve (1) using computer software. Using the *Maple* dsolve command with the abserr option, see if you can obtain accurate results at the points $x = 1, 4, 8, 12$ listed in Table 1.

3. One can see if a computed solution exhibits instability, as did the solution obtained by the midpoint rule and plotted in Fig. 2, when we have the exact solution to compare it with. In practice, of course, we don't have the exact solution to compare with; if we did, then we would not be solving numerically in the first place. Thus, when a computed solution exhibits an oscillatory behavior how do we know that it is incorrect; perhaps the exact solution has exactly that oscillatory behavior? One way to find out is to rerun the solution with h halved. If the oscillatory behavior is part of the exact solution, then the new results will oscillate every two steps rather than every step. Using this idea, run the case shown in Fig. 2 twice, for $h = 0.05$ and $h = 0.025$, and comment on whether the results indicate a true instability or not.

4. We derived the solution (12) of the difference equation (7) in the text. Verify, by direct substitution, that (12) does satisfy (7) for any choice of the arbitrary constants C_1 and C_2.

5. In (13) we showed that $\left(Ah + \sqrt{1 + A^2h^2}\right)^n \sim e^{Ax_n}$ as $h \to 0$, yet it would appear that $\left(Ah + \sqrt{1 + A^2h^2}\right)^n \sim \left(\sqrt{1}\right)^n = 1$. Explain the apparent contradiction.

6. The purpose of this exercise is to explore the validity of the rule of thumb that we gave regarding the solution of the equation $y' = f(x, y)$ by the midpoint rule – namely, that the method should be stable if $\partial f/\partial y > 0$ and unstable if $\partial f/\partial y < 0$ over the region of interest. Specifically, in each case apply the rule of thumb and draw what conclusions you can about the stability of the midpoint rule solution of the given problem. Then, program and run the midpoint rule with $h = 0.05$, say, over the given x interval. Discuss the numerical results and whether the rule of thumb was correct. (Since the midpoint rule is not self-starting, use ten Euler steps from $x = 0$ to $x = 0.05$ to get the method started.)

(a) $y' = e^{x+y};\quad y(0) = 1, 0 \leq x \leq 4$
(b) $y' = e^{x-y};\quad y(0) = 1, 0 \leq x \leq 4$
(c) $y' = (4 - x)y;\quad y(0) = 1, 0 \leq x \leq 10$
(d) $y' = (x - 1)y;\quad y(0) = 1, 0 \leq x \leq 5$
(e) $y' = (x + y)/(1 + x);\quad y(0) = 2, 0 \leq x \leq 4$

7. We stated in the text that the results in Table 2 are consistent with a critical h value of 0.002 because the calculations change from unstable to stable as h decreases from 0.01 to 0.001. Program and carry out the Euler calculation of the solution to the initial-value problem (18) using $h = 0.0021$ and 0.0019, out to around $x = 1$, and see if these h values continue to bracket the change from unstable to stable. (You may try to bracket h_{cr} even more tightly if you wish.)

8. (*Stability of second-order Runge–Kutta methods*) In Section 6.3 we derived the general second-order Runge–Kutta method, which includes these as special cases: the *improved Euler method*,

$$y_{n+1} = y_n + \frac{h}{2}\left\{f(x_n, y_n) + f\left[x_{n+1}, y_n + hf(x_n, y_n)\right]\right\}, \tag{8.1}$$

and the *modified Euler method*,

$$y_{n+1} = y_n + hf[x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n)]. \tag{8.2}$$

(a) For the test equation $y' = Ay$, show that the improved Euler method is strongly stable for sufficiently small h (i.e., as $h \to 0$). For the case where $A < 0$, show that that stability is achieved only if $h < h_{cr} = 2/|A|$.
(b) For the test equation $y' = Ay$, show that the modified Euler method is strongly stable for sufficiently small h (i.e., as $h \to 0$). For the case where $A < 0$, show that that stability is achieved only if $h < h_{cr} = 2/|A|$.

9. (*Strong stability of the multi-step Adams–Bashforth method*) Recall, from Section 6.3, the fourth-order Adams–Bashforth method

$$y_{n+1} = y_n + \frac{h}{24}\left(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}\right), \tag{9.1}$$

where $f_n = f(x_n, y_n)$. This exercise is to show that the "AB" method is strongly stable for sufficiently small h (i.e., as $h \to 0$) even though it is a multi-step method.

(a) Consider the test equation $y' = Ay$, where the constant A can be positive or negative; that is, let $f(x, y) = Ay$ be a solution of the fourth-order difference equation (9.1) in the form $y_n = \rho^n$, show that ρ must satisfy the fourth-degree characteristic equation

$$\rho^4 - (1 + 55\alpha)\,\rho^3 + 59\alpha\rho^2 - 37\alpha\rho + 9\alpha = 0, \tag{9.2}$$

where $\alpha = Ah/24$.
(b) Notice that as h tends to zero so does α, and (9.2) reduces to $\rho^4 - \rho^3 = 0$, with the roots $= 0, 0, 0, 1$. Thus, if we denote the roots of (9.2) as $\rho_1, \ldots, \rho_4$, then we see that the first three of these tend to zero and the last to unity as $h \to 0$, and the general solution for y_n behaves as

$$y_n = C_1\rho_1^n + C_2\rho_2^n + C_3\rho_3^n + C_4\rho_4^n \sim C_4 1^n \tag{9.3}$$

as $h \to 0$. Since p^n tends to zero, unity, or infinity, depending upon whether p is smaller than, equal to, or greater than unity, we need to examine the 1^n term in (9.3) more closely. Specifically, seeking ρ_4 in the power series form $\rho_4 = 1 + a\alpha + \cdots$, put that form into (9.2). Equating coefficients of α on both sides of that equation through first-order terms, show that $a = 24$. Thus, in place of (9.3) we have the more informative statement

$$y_n \sim C_4(1 + 24\alpha)^n = C_4(1 + Ah)^n \sim C_4 e^{Ax_n} \tag{9.4}$$

as $n \to \infty$. Show why the final step in (9.4) is true. Since the right-hand side of (9.4) is identical to the exact solution $y(x)$ of the given differential equation, whether A is positive or negative, we conclude that the AB method is strongly stable for sufficiently small h.

10. (*Strong stability of the multi-step Adams–Moulton method*) Recall, from Section 6.3, the fourth-order Adams–Moulton method,

$$y_{n+1} = y_n + \frac{h}{24}\left(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}\right). \tag{10.1}$$

Proceeding along the same lines as outlined in Exercise 9, show that the AM method is stongly stable for sufficiently small h.

11. Derive the general solution of the given difference equation. If initial conditions are specified, then find the corresponding particular solution. In each case $n = 0, 1, 2, \ldots$.

(a) $y_{n+1} - 4y_n = 0$; $y_0 = 5$
(b) $y_{n+2} - y_n = 0$; $y_0 = 1$, $y_1 = 3$
(c) $y_{n+2} + y_{n+1} - 6y_n = 0$; $y_0 = 9$, $y_1 = -2$
(d) $y_{n+2} - 4y_{n+1} + 3y_n = 0$; $y_0 = 3$, $y_1 = 1$
(e) $y_{n+2} + 3y_{n+1} + 2y_n = 0$; $y_0 = 5$, $y_1 = 7$
(f) $y_{n+3} - y_{n+2} - 4y_{n+1} + 4y_n = 0$; $y_0 = 3$, $y_1 = 5$, $y_2 = 9$
(g) $y_{n+4} - 5y_{n+2} + 4y_n = 0$
(h) $y_{n+4} - 6y_{n+2} + 8y_n = 0$

12. (a)–(h) Use computer software to obtain the general solution of the corresponding problem in Exercise 11, and the particular solution as well, if initial conditions are given.

13. (*Repeated roots*) Recall from the theory of linear homogeneous differential equations with constant coefficients that if, when we seek $y(x) = e^{\lambda x}$, λ is a root of multiplicity k of the characteristic equation, then it gives rise to the solutions $y(x) = \left(C_1 + C_2x + \cdots + C_kx^{k-1}\right)e^{\lambda x}$, where $C_1, \ldots, C_k$ are arbitrary constants. An analogous result holds for difference equations. Specifically, verify that the characteristic equation of $y_{n+2} - 2by_{n+1} + b^2y_n = 0$ has the root b with multiplicity 2, and that $y_n = (C_1 + C_2n)b^n$.

14. Show that if $y_n^{(1)}$ and $y_n^{(2)}$ are solutions of the second-order linear homogeneous difference equation $a_0(n)y_{n+2} + a_1(n)y_{n+1} + a_2(n)y_n = 0$, and Y_n is any particular solution of the nonhomogeneous equation $a_0(n)y_{n+2} + a_1(n)y_{n+1} + a_2(n)y_n = f_n$, then $y_n = C_1 y_n^{(1)} + C_2 y_n^{(2)} + Y_n$ is a solution of the nonhomogeneous equation.

15. (*Nonhomogeneous difference equations*) For the given equation, first find the general solution of the homogeneous equation. Then adapt the method of undetermined coefficients from the theory of differential equations and find a particular solution. Finally, give the general solution of the given nonhomogeneous difference equation. (First, read Exercise 14.)

(a) $y_{n+1} - 3y_n = n$
(b) $y_{n+1} - 2y_n = 3\sin n$
(c) $y_{n+1} - y_n = 2 + \cos n$
(d) $y_{n+2} - 5y_{n+1} + 6y_n = 2n^2 - 6n - 1$
(e) $y_{n+2} + y_{n+1} - 2y_n = n^2$
(f) $y_{n+2} - 4y_n = 6n^2 - 1$
(g) $y_{n+2} - y_n = e^n$
(h) $y_{n+4} - 7y_{n+2} + 12y_n = n + 6$

16. (a) – (h) Use computer software to obtain the general solution of the corresponding problem in Exercise 15.

Chapter 6 Review

To solve a differential equation $y' = f(x, y)$ numerically, one begins by discretizing the problem and working with the discrete variables x_n, y_n in place of the continuous variables $x, y(x)$. The solution is accomplished using a numerical algorithm that gives y_{n+1} in terms of y_n in the case of a single-step method, or in terms of y_n and one or more of the preceding values $y_{n-1}, y_{n-2}, \ldots$ in the case of a multi-step method. If the algorithm gives y_{n+1} explicitly, then it is said to be of **open** type; if not it is of **closed** type. All methods considered in this chapter are of open type, except for the Adams–Moulton method (Section 6.3.5).

Decomposing the error as

$$\text{Total error } = y(x_n) - y_n^* = [y(x_n) - y_n] + [y_n - y_n^*],$$

where $y(x_n)$ is the exact solution of the differential equation, y_n is the exact solution of the numerical algorithm (i.e., carried out on a perfect machine, having no roundoff error), and y_n^* is the actual solution of the algorithm (carried out on a real machine), we call $y(x_n) - y_n$ the **accumulated truncation error** and $y_n - y_n^*$ the **accumulated roundoff error**. If the former is of order $O(h^p)$ at a fixed x point as $h \to 0$, then the method is said to be of **order** p. If $p > 0$, then the accumulated truncation error tends to zero and the method is described as **convergent**. The greater p is, the more accurate the method is for a given step size h. Besides requiring of a method that it be convergent, we also require that it be **stable**; that is, we require that the accumulated roundoff error (which is inevitable in a real machine) remain small compared to the exact solution of the differential equation.

We begin with the **Euler method** $y_{n+1} = y_n + f(x_n, y_n)h$, which is simple but not very accurate because it is only a first-order method. Desiring greater accuracy, we introduce the second-order **midpoint rule** and second- and fourth-order **Runge–Kutta methods**, the latter providing us with an accurate general-purpose differential equation solver.

One's interest in higher-order methods is not just a matter of accuracy because, in principle, one could rely exclusively on the simple and easily programmed Euler method, and make h small enough to achieve any desired accuracy. There are two problems with that idea. First, as h is decreased the number of steps increases and one can expect the numerical roundoff error to grow, so that it may not be possible to achieve the desired accuracy. Second, there is the question of economy. For instance, while the fourth-order Runge–Kutta method (for example) is about four times as slow as the Euler method (because it requires four function evaluations per step compared to one for the Euler method), the gain in accuracy that it affords is so great that we can use a step size much more than four times that needed by the Euler method for the same accuracy, thereby resulting in greater economy.

Naturally, higher-order methods are more complex and hence more tedious to program. Thus, we strongly urge (in Section 6.3.4) the empirical estimation of the order, if only as a check on the programming and implementation of the method.

In Section 6.4 we showed that the methods developed for the single equation $y' = f(x, y)$ can be used to solve systems of equations and higher-order equations as well. There we also study boundary-value problems, and find them to be significantly more difficult than initial-value problems. However, we show how to use the principle of superposition to convert a boundary-value problem to one or more problems of initial-value type, provided that the problem is linear.

Finally, in Section 6.5 we look at "what can go wrong," mostly insofar as numerical instability due to the growth of roundoff error, and an analytical approach is put forward for predicting whether a given method is stable. Actually, stability depends not only on the solution algorithm but also on the differential equation, and our analyses are for the simple test equation $y' = Ay$ rather than for the general case $y' = f(x, y)$. We find that whereas the differential equation $y' = Ay$ is of first order, the difference equation expressed by the algorithm is of higher order if the method is of multi-step type. Thus, it has among its solutions the exact solution (as $h \to 0$) and one or more extraneous solutions as well. It is those extraneous solutions that can cause instability. For instance, the midpoint rule is found to be stable if $A > 0$ and unstable if $A < 0$; we classify it as weakly stable because its stability depends upon the sign of A. However, the fourth-order Adams–Bashforth and Adams–Moulton methods are stable, even though they are multistep methods because the extraneous solutions do not grow. Single-step methods such as Euler and those of Runge–Kutta type do not give rise to extraneous solutions and are stable.

Finally, we stress that even if a method is stable as $h \to 0$, h needs to be reduced below some critical value for that stability to be manifested.

Chapter 7

Qualitative Methods: Phase Plane and Nonlinear Differential Equations

7.1 Introduction

This is the final chapter on ordinary differential equations, although we do return to the subject in Chapter 11, where we reconsider systems of linear differential equations using matrix methods.

Interest in nonlinear differential equations is virtually as old as the subject of differential equations itself, which dates back to Newton, but little progress was made until the late 1880's when the great mathematician and astronomer *Henri Poincaré* (1854 – 1912) took up a systematic study of the subject in connection with celestial mechanics. Realizing that nonlinear equations are rarely solvable analytically, and not yet having the benefit of computers to generate solutions numerically, he sidestepped the search for solutions altogether and instead sought to answer fundamental questions about the qualitative and topological nature of solutions of nonlinear differential equations without actually finding them.

The entire chapter reflects either his methods, such as the use of the so-called "phase plane" and focusing attention upon the "singular points" of the equation, or the spirit of his approach. In addition, however, we can now rely heavily upon computer simulation. Thus, our approach in this chapter is a blend of a qualitative, topological, and geometric approach, with quantitative results obtained readily with computer software.

Though Poincaré's work was motivated primarily by problems of celestial mechanics, the subject began to attract broader attention during and following World War II, especially in connection with nonlinear control theory. In the postwar years, interest was stimulated further by the publication in English of N. Minorsky's *Nonlinear Mechanics* (Ann Arbor, MI: J. W. Edwards) in 1947. With that and other books, such as A. Andronov and C. Chaikin's *Theory of Oscillations* (Princeton:

Princeton University Press, 1949) and J. J. Stoker's *Nonlinear Vibrations* (New York: Interscience, 1950) available as texts, the subject appeared in university curricula by the end of the 1950's. With that base, and the availability of digital computers by then, the subject of nonlinear dynamics, generally known now as *dynamical systems*, has blossomed into one of the most active research areas, with applications well beyond celestial mechanics and engineering – to biological systems, the social sciences, economics, and chemistry. The shift from the orderly determinism of Newton to the often nondeterministic chaotic world of Mitchell Feigenbaum, E. N. Lorenz, Benoit Mandelbrot, and Stephen Smale has been profound. For a wonderful historical discussion of these changes we suggest the little book *Mathematics and the Unexpected* by Ivar Ekeland (Chicago: University of Chicago Press, 1988).

7.2 The Phase Plane

To introduce the phase plane, consider the system

$$mx'' + kx = 0 \tag{1}$$

governing the free oscillation of the simple harmonic mechanical oscillator shown in Fig. 1. Of course we can readily solve (1) and obtain the general solution $x(t) = C_1 \cos \omega t + C_2 \sin \omega t$, where $\omega = \sqrt{k/m}$ is the natural frequency – or, equivalently,

$$x(t) = A \sin(\omega t + \phi), \tag{2}$$

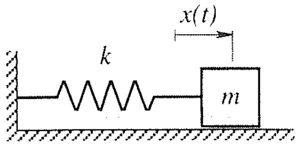


Figure 1. Simple harmonic mechanical oscillator.

where A and ϕ are the amplitude and phase angle, respectively. To present this result graphically, one can plot x versus t and obtain any number of sine waves of different amplitude and phase, but let us proceed differently.

We begin by re-expressing (1), equivalently, as the system of first-order equations

$$\frac{dx}{dt} = y, \tag{3a}$$

$$\frac{dy}{dt} = -\frac{k}{m}x, \tag{3b}$$

as is discussed in Section 3.9. The auxiliary variable y, defined by (3a), happens to have an important physical significance, it is the velocity, but having such significance; is not necessary. Next, we deviate from the ideas presented in Section 3.9 and divide (3b) by (3a), obtaining

$$\frac{dy}{dx} = -\frac{k}{m}\frac{x}{y} \qquad \text{or} \qquad my\,dy + kx\,dx = 0, \tag{4}$$

integration of which gives

$$\frac{1}{2}my^2 + \frac{1}{2}kx^2 = C. \tag{5}$$

Since $y = dx/dt$, (5) is a first-order differential equation. We could solve for y (i.e., dx/dt), separate variables, integrate again, and eventually arrive at (2) once again. Instead, let us take (5) as our end result and plot the one-parameter family of ellipses that it defines (Fig. 2), the parameter being the integration constant C. In this example C happens to the total energy (kinetic energy of the mass plus potential energy of the spring); $C = 0$ gives the "point ellipse" $x = y = 0$ and the greater the value of C, the larger the ellipse.

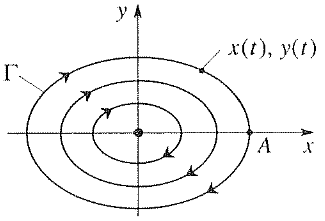


Figure 2. Phase portrait of (1).

It is customary to speak of the x, y plane as the **phase plane**. Each integral curve represents a possible motion of the mass, and each point on a given curve represents an instantaneous state of the mass (the horizontal coordinate being the displacement and the vertical coordinate being the velocity). Observe that the time t enters only as a parameter, through the parametric representation $x = x(t)$, $y = y(t)$. So we can visualize the representative point $x(t), y(t)$ as moving along a given curve as suggested by the arrows in Fig. 2. The direction of the arrows is implied by the fact that $y = dx/dt$, so that $y > 0$ implies that $x(t)$ is increasing and $y < 0$ implies that $x(t)$ is decreasing. One generally calls the integral curves **phase trajectories**, or simply **trajectories**, to suggest the idea of movement of the representative point. A display of a number of such trajectories in the phase plane is called a **phase portrait** of the original differential equation, in this case (1). Of course, there is a trajectory through each point of the phase plane, so if we showed *all* possible trajectories we would simply have a black picture; the idea is to plot enough trajectories to establish the key features of the phase portrait.

What are the advantages of presenting results in the form of a phase portrait, rather than as traditional plots of $x(t)$ versus t? One advantage of the phase portrait is that it requires only a "first integral" of the original second-order equation such as equation (5) in this example, and sometimes we can obtain the first integral even when the original differential equation is nonlinear. For instance, let us complicate (1) by supposing that the spring force is not given by the linear function $F_s = kx$, but by the nonlinear function $F_s = ax + bx^3$, and suppose that $a > 0$ and $b > 0$ so that the spring is a "hard" spring: it grows stiffer as x increases (Fig. 3), as does a typical rubber band. If we take $a = b = m$, say, for definiteness and for simplicity, then in place of (1) we have the nonlinear equation

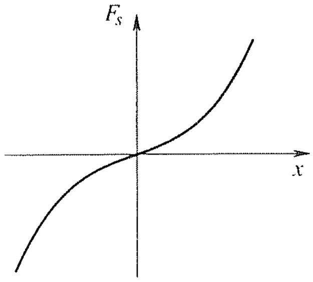


Figure 3. Hard spring.

$$x'' + x + x^3 = 0. \tag{6}$$

Proceeding as before, we re-express (6) as the system

$$x' = y, \tag{7a}$$
$$y' = -x - x^3. \tag{7b}$$

Division gives

$$\frac{dy}{dx} = -\frac{x + x^3}{y} \qquad \text{or} \qquad y\,dy + (x + x^3)dx = 0, \tag{8}$$

which yields the first integral

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 + \frac{1}{4}x^4 = C. \tag{9}$$

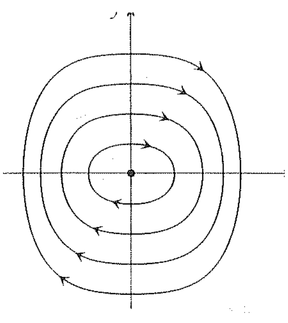


Figure 4. Phase portrait of (6) (hard spring).

In principle, if we plot these curves for various values of C we can obtain the phase portrait shown in Fig. 4. More conveniently, we generated the figure by using the *Maple* phaseportrait command discussed at the end of this section. A comparable phase portrait plotting capability is provided in numerous other computer software systems.

To repeat, one advantage of the phase portrait presentation is that it requires only a first integral. In the present case (6) was nonlinear due to the x^3 term, yet its first integral (9) was readily obtained.

A second attractive feature of the phase portrait is its compactness. For instance, observe that the single phase trajectory Γ in Fig. 2 corresponds to an entire family of oscillations of amplitude A, several of which are shown in Fig. 5 since any point on Γ can be designated as the initial point ($t = 0$): if the initial point on Γ is $(A, 0)$, then we get the curve #1 in Fig. 5; if the initial point on Γ is a bit counterclockwise of $(A, 0)$ then we get the curve #2; and so on. Passing from the x, t plane to the x, y plane, the infinite family of curves shown in Fig. 5 collapse onto the single trajectory Γ in Fig. 2. Put differently, whereas the solution (2) of equation (1) is a two-parameter family of curves in x, t space (the parameters being A and ϕ), (5) is only a one-parameter family of curves in the x, y plane (the parameter being C). That compactness can be traced to the division of (3b) by (3a) or (7b) by (7a) for that step essentially eliminates the time t.

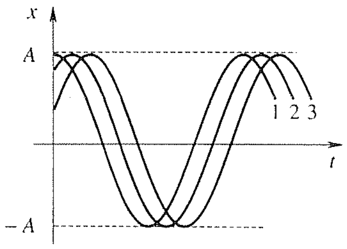


Figure 5. Solutions $x(t)$ corresponding to the trajectory Γ.

To learn about nonlinear systems, it is useful to contrast the phase portraits of the linear oscillator governed by (1) and the nonlinear oscillator governed by (6), and given in Figs. 2 and 4, respectively. The phase portrait in Fig. 2 is extremely simple in the sense that all the trajectories are geometrically similar, differing only in scale. That is, if a trajectory is given by $x = X(t)$, $y = Y(t)$, then $x = \kappa X(t)$, $y = \kappa Y(t)$ is also a trajectory for every possible scale factor κ, be it positive, negative, or zero. That result holds not only for the system (3) but for any constant-coefficient linear homogeneous system

$$\begin{aligned} x' &= ax + by, \\ y' &= cx + dy. \end{aligned} \tag{10}$$

In contrast, consider the phase portrait of the nonlinear equation $x'' + \alpha x + \beta x^3 = 0$ shown in Fig. 4. In that case the trajectories are not mere scalings of each other; there is distortion of shape from one to another, and that distortion is due entirely to the nonlinearity of the differential equation. The innermost trajectories approach ellipses [because as smaller and smaller motions are considered the x^4 becomes more and more negligible compared to the other terms in (9)], and the outer ones become more and more distorted as the effect of the x^4 term grows in (9).

Thus, whereas the phase portrait of the linear equation (1) amounts to a single kind of trajectory, repeated endlessly through scalings, that of the nonlinear equation (6) is made up of an infinity of different kinds of trajectories. That richness is a hallmark of nonlinear equations, as we shall see in the next example and in the sections that follow.

Before turning to the next example, let us complement the phase portrait in Fig. 4 with representative plots of $x(t)$ versus t. We choose the two sets of initial conditions: $x(0) = 0.5$, $x'(0) = 0$ and $x(0) = 1$, $x'(0) = 0$. The results are shown in Fig. 6, together with the corresponding solutions of the linear equation $x'' + x = 0$ (shown as dotted) for reference. Besides the expected distortion we also observe that the frequency of the oscillation is amplitude dependent for the nonlinear case: the frequency increases as the amplitude increases. In contrast, for the linear equation (1) the frequency $\omega = \sqrt{k/m} = 1$ is a constant, independent of the amplitude.

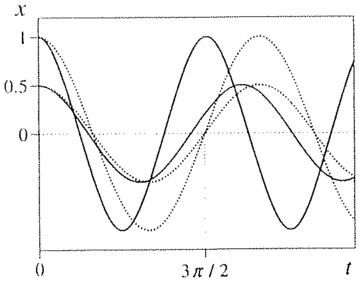


Figure 6. Effects of nonlinearity on x(t).

Above, we mentioned the richness of the sets of solutions to nonlinear differential equations. A much more striking example of that richness is obtained if we reconsider the nonlinear oscillator, this time with a "soft" spring – that is, with $F_s = ax - bx^3$ ($a > 0$ and $b > 0$) as sketched in Fig. 7. Again setting $a = b = m$ we have, in place of (6),

$$x'' + x - x^3 = 0. \tag{11}$$

In place of (9) we have

$$\frac{1}{2}y^2 + \frac{1}{2}x^2 - \frac{1}{4}x^4 = C, \tag{12}$$

and in place of the phase portrait shown in Fig. 4 we obtain the strikingly different one shown in Fig. 8. We continue to study this example in Section 7.3, but even now we can make numerous interesting observations. First, whereas all of the motions revealed in the phase plane for the hard spring (Fig. 4) are qualitatively similar oscillatory motions – give or take some distortion from one to another we see in Fig. 8 a number of qualitatively different types of motion, and these are

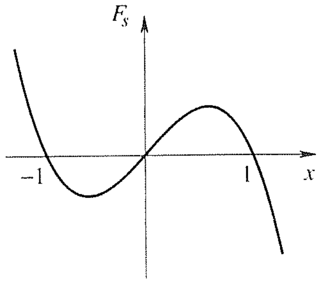


Figure 7. Soft spring.

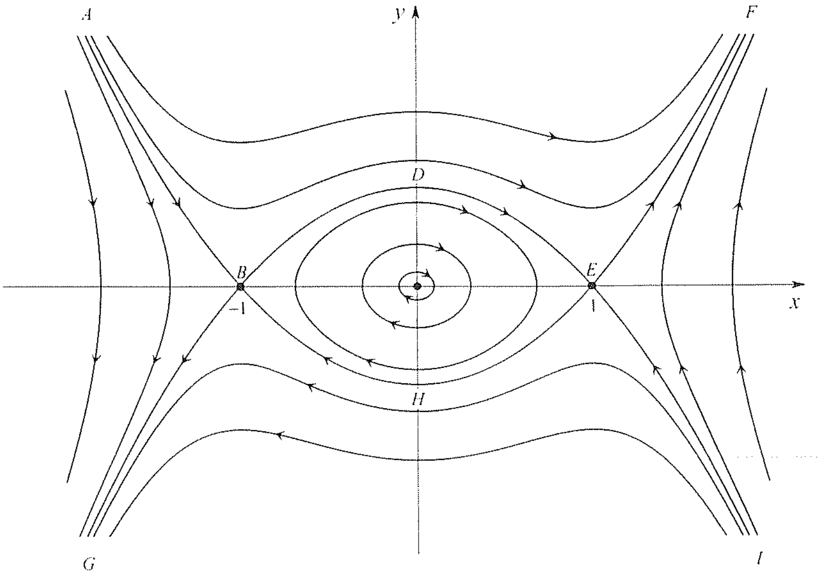


Figure 8. Phase portrait of (11) (soft spring).

separated by the trajectories $ABHEF$ and $GBDEI$ which, together, are called a **separatrix**.

Before examining these different types of motion let us distinguish between the physical velocity $dx/dt = x' = y$ of the mass and the **phase velocity**

$$\frac{ds}{dt} = s' = \sqrt{x'^2 + y'^2} \tag{13}$$

of the representative point $P = (x(t), y(t))$ in the phase plane as it moves along a trajectory, where s is arc length along that trajectory. Observe from (13) that $s' = 0$, and the representative point P is stationary if and only if both $x' = 0$ and $y' = 0$ at that point. If $x' = 0$ and $y' = 0$ at a point $P_0 = (x_0, y_0)$, then we call P_0 an **equilibrium point**, or **fixed point**, because then $x(t) = x_0$ and $y(t) = y_0$ remain constant; if we start at that point we remain at that point for all $t > 0$.

Let us identify the equilibrium points (if any) for the present soft-spring example. Solving

$$x' = y = 0, \tag{14a}$$
$$y' = -x + x^3 = 0 \tag{14b}$$

gives $y = 0$ and $x = 0, \pm 1$, so there are three equilibrium points:

$$(-1, 0),\ (0, 0),\ (1, 0), \tag{15}$$

which correspond to B, the origin, and E in Fig. 8. Physically, $x = 0$ and ± 1 are the three x values for which the spring force F_s is zero (Fig. 7).

Consider the motions inside the $BDEHB$ "football." First, each closed loop trajectory Γ inside $BDEHB$ corresponds to a periodic solution $x(t)$ (likewise for each closed loop trajectory in Figs. 2 and 4), and is therefore called a **closed orbit**. [Recall from Section 5.7 that $x = x(t)$ is periodic with period T if $x(t+T) = x(t)$ for all $t \geq 0$.] Actually, how can we be sure that each Γ represents a periodic motion? Well, the representative point does trace out the given trajectory Γ over and over; it cannot stop because there are no equilibrium points on Γ. Further, the phase speed s' at any point (x, y) on Γ is the same, from one traversal to the next, because

$$s' = \sqrt{x'^2 + y'^2} = \sqrt{y^2 + (x - x^3)^2} \tag{16}$$

is a function only of x and y, not t. Thus, as P traverses Γ, over and over, each time in the same manner, the projection $x(t)$ of P onto the x axis generates a periodic function $x(t)$.

Next, consider part of the separatrix itself. To obtain the equation of the separatrix we use (12) and require the curve to go through the point E (or, equivalently, B) – namely, $x = 1$ and $y = 0$. That step gives $C = 1/4$, so $ABHEF$ and $GBDEI$ are given by

$$y = \pm\sqrt{\frac{1}{2} - x^2 + \frac{1}{2}x^4}, \tag{17}$$

respectively. Beginning at D, say, the representative point P moves rightward on DE and approaches the equilibrium point E. Does it reach E in finite time and then remain there, or does it approach E asymptotically as $t \to \infty$? To answer that question we use (16). Let the tangent line to the curve DEI, at E, be $y = m(x-1)$. [We could determine the slope m by differentiating (17), but the value of m will not be important.] Then, since $ds = \sqrt{1+(dy/dx)^2}\,dx \sim \sqrt{1+m^2}\,dx$ as $P \to E$, we can replace s' in (16) by $\sqrt{1+m^2}\,dx/dt$, and y by $m(x-1)$, so (16) becomes

$$\begin{aligned}\sqrt{1+m^2}\,\frac{dx}{dt} &\sim \sqrt{m^2(x-1)^2 + x^2(x+1)^2(x-1)^2}\\ &\sim -\sqrt{m^2+4}\,(x-1),\end{aligned} \tag{18}$$

where the negative square root has been chosen since $dx/dt > 0$ as $P \to E$ on DE, whereas $x - 1 < 0$ on DE, and where the last step in (18) is left for the exercises. The upshot is that

$$\frac{dx}{dt} \sim \gamma(1-x) \tag{19}$$

as $P \to E$, for some finite positive constant γ. Thus,

$$\frac{dx}{1-x} \sim \gamma\,dt \tag{20}$$

so

$$-\ln{(1-x)} \sim \gamma\,t + \text{constant}, \tag{21}$$

and we can now see that $t \to \infty$ as $P \to E$ (i.e., as $x \to 1$). Thus, P does not reach E in finite time but only asymptotically as $t \to \infty$. Similarly, if we begin at point H and go backward in time, then we reach E only as $t \to -\infty$.

Let us return now to the region inside of the football and consider any closed orbit Γ. As the size of Γ shrinks to zero, Γ tends to the elliptical (actually circular because of our choice $a = m$) shape $y^2 + x^2 =$ constant, and the period of the motion tends to 2π [since the solution of the linearized problem is $x(t) = A\sin{(t+\phi)}$]. At the other extreme, as Γ gets larger it approaches the pointed shape $BDEHB$. Bearing in mind that it takes infinite time to reach E along an approach from D, it seems evident that the period of the Γ motion must tend to infinity as Γ approaches $BDEHB$; we will ask you to explore this point in the exercises. From a physical point of view, the idea is that not only is the "flow" zero *at* E (where $x' = y' = 0$), it is very slow in the neighborhood of E. If Γ is any closed trajectory that is just barely inside of $BDEHB$, then part of Γ falls within that stagnant neighborhood of E (similarly at B). The representative point P moves very slowly there, hence the period is very large. We reiterate that although each closed loop inside the football corresponds to a periodic motion, the closed loop $BDEHB$ does not. In fact, although BDE and EHB meet at B and E they are distinct trajectories. On BDE, t varies from $-\infty$ at B to $+\infty$ at E; likewise, on EHB t varies from $-\infty$ at E to $+\infty$ at B. Thus, if we begin on BDE we can never get to EHB, and vice versa.

Finally, it should be evident that every trajectory that is *not* within the football corresponds to a nonperiodic motion. Thus, BDE and EHB form a transition from nonperiodic to periodic motions.

Thus far we have studied the three differential equations (1), (6), and (11). In each case we have changed the single second-order differential equation to a system of two first-order equations by setting $x' = y$ and then studied them in the x, y phase plane. More generally, we consider in this chapter systems of the form

$$x' = P(x, y), \tag{22a}$$
$$y' = Q(x, y). \tag{22b}$$

That is, $P(x, y)$ need not equal y, and the system need not be a restatement of a single second-order equation. Rather, it might arise directly in the form (22).

For instance, suppose that two species of fish coexist in a lake, say bluegills and bass. The bluegills, with population x, feed on vegetation which is available in unlimited quantity, and the bass, with population y, feed exclusively on the bluegills. If the two species were separated, their populations could be assumed to be governed approximately by the rate equations

$$x' = \alpha x, \qquad y' = -\beta y, \tag{23}$$

where the populations $x(t)$ and $y(t)$ are considered to be large enough so that they can be approximated as continuous rather than discrete (integer valued) variables, and α, β are (presumably positive) constants that reflect net birth/death rates. The species are not separated, however, so we expect the effective α to decrease as y increases and the effective β to decrease as x increases. An approximate revised model might then be expressed as

$$x' = (\alpha - \gamma y)x, \tag{24a}$$
$$y' = -(\beta - \delta x)y, \tag{24b}$$

which system is indeed of the form (22). This ecological problem is well known as *Volterra's problem*, and we shall return to it later.

The system (22) is said to be **autonomous** because there is no explicit dependence on the independent variable (the time t here but which could have some other physical or nonphysical significance). Surely not all systems are autonomous, but that class covers a great many cases of important interest, and that is the class that is considered in phase plane analysis and in this chapter. Because (22a,b) are autonomous, any explicit reference to t (namely, the t derivatives) can be suppressed by dividing one equation by the other and obtaining

$$\frac{dy}{dx} = \frac{P(x, y)}{Q(x, y)}, \tag{25}$$

where we now change our point of view and regard y as a function of x in the x, y phase plane, rather than x and y as functions of t. If the system (22) were not autonomous; that is, if it were of the form $x' = P(x, y, t)$ and $y' = Q(x, y, t)$, one could still make it autonomous by re-expressing it, equivalently, as

$$\begin{aligned} x' &= P(x, y, z), \\ y' &= Q(x, y, z), \\ z' &= 1, \end{aligned}$$

but then the trajectories exist in the three-dimensional x, y, z space, and that case is more complicated. In this chapter we continue to consider the autonomous case (22) and the two-dimensional x, y phase plane.

Closure. As explained immediately above, our program in this section is to show the advantages of recasting an autonomous system (22) (which could, but need not, arise from a single second- order equation by letting x' be an auxiliary dependent variable y) in the form (25) and then study the solutions of that equation in the two-dimensional x, y phase plane. One advantage is that (25) can sometimes be solved analytically even if (22a,b) are nonlinear. Indeed, our primary interest in Chapter 7 is in the nonlinear case. We find that the phase portrait provides a remarkable overview of the system dynamics, and the hard- and soft-spring oscillator examples begin to reveal some of the phenomenological richness of nonlinear systems. We do not suggest the use of the phase plane as a substitute for obtaining and plotting solutions of (22) in the more usual way, x versus t and y versus t. Rather, we suggest that to understand a complex nonlinear system one needs to combine several approaches, and for autonomous systems the phase plane is one of the most valuable. Finally, we ask you to observe how the phase plane discussion is more qualitative and topological than lines of approach developed in the preceding chapters. For instance, regarding Fig. 8 we distinguish the qualitatively different types of motion such as the periodic orbits within $BDEHB$, the transitional motions on the separatrix itself, and the nonperiodic motions as well.

We distinguish between the physical velocity $x'(t)$ of the mass, in the preceding examples, and the phase velocity $s'(t)$, which is the velocity of the representative point $x(t), y(t)$ in the phase plane. It is useful, conceptually, to think of the $x'(t), y'(t)$ velocity field as the velocity field of a "flow" such as a fluid flow in the phase plane.

Finally, we mention that in the hard- and soft-spring oscillators, (6) and (8), we meet special cases of the extremely important **Duffing equation**, to which we return in a later section.

Computer software. Here is how we generate the phase portrait shown in Fig. 8 using the *Maple* **phaseportrait** command. First, enter

with(DEtools):

and return, to gain access to the phaseportrait command. Note the colon, whereas *Maple* commands are followed by semicolons. Next, enter

phaseportrait($[y, -x + x\hat{\ }3]$, $[t, x, y]$, $t = -20..20$, $\{[0, 0, 0.1], [0, 0, 0.3]$,

$[0,0,0.6],[0,0,0.70710781],[0,0,-0.70710781],[0,0,0.9],[0,0,-0.9],$
$[0,0,1.25],[0,0,-1.25],[0,1.5,0.8838834761],[0,1.5,-0.8838834761],$
$[0,1.4,0],[0,1.8,0],[0,-1.5,0.8838834761],[0,-1.5,-0.8838834761],$
$[0,-1.4,0],[0,-1.8,0]\}$, stepsize $= 0.05$, $y = -1.8..1.8$, $x = -2..2$,
scene $= [x, y]$);

and return. In $[y, -x + x\hat{}3]$ the items are the right-hand sides of the first and second differential equations, respectively; $[t, x, y]$ are the independent variable and dependent variables; $t = -20..20$ is the range of integration of the differential equations; within { } are the initial t, x, y points chosen in order to generate the trajectories shown in Fig. 8. After those points the remaining items are optional: stepsize= 0.05 sets the stepsize h in the Runge–Kutta–Fehlberg integration because the default value would be (final t – initial $t)/20 = (20 + 20)/20 = 2$, which would give too coarse a plot (as found by experience); $y = -1.8..1.8, x = -2..2$ gives a limit to the x, y region, with the t integrations terminated once a trajectory leaves that region; scene $= [x, y]$ specifies the plot to be a two-dimensional plot in the x, y plane, the default being to give a three-dimensional plot in t, x, y space. There are additional options that we have not used, one especially useful option for phase plane work being the arrow option, which gives a lineal element grid. The elements can be barbed or not. To include thin barbed arrows, type a comma and then arrows=THIN after the last option. Thus, we would have ... scene $= [x, y]$, arrows =THIN);. In place of THIN type SLIM, or THICK for thicker arrows. For unbarbed lineal elements, use LINE in place of these. The order of the options is immaterial.

Observe that the separatrix must be generated separately as BDE, EHB, AB, GB, EF, and IE. To generate BDE, for instance, we determine the coordinates of D. The equation of the entire separatrix is given by (12), where C is determined by using the x, y pair $1, 0$ (namely, the point E, which is a known point on the separatrix). Thus, putting $x = 1$ and $y = 0$ into the left side of (12) gives $C = 1/4$. Next, put $x = 0$ and solve for y, obtaining $y = 1/\sqrt{2} = 0.770710781$ at D. Then, with D as the initial point we need t to go from $-\infty$ to $+\infty$ to generate BDE. By trial, we find that -20 to $+20$ suffices for this segment and all others; similarly, we generate EF by using a point on EF at $x = 1.5$, and determine y at that point from the separatrix equation. That calculation gives $y = 0.8838834761$.

Notice that to generate Fig. 8 with phaseportrait we need to already know something about the phaseportrait – the equation of the separatrix, (12), so that we can choose suitable initial points on AB, GB, BDE, EHB, EF, and IE.

Suppose that we desire only the lineal element field, over $0 < x < 4$ and $0 < y < 4$, say. We can get it from phaseportrait as follows:

phaseportrait($[y, -x + x\hat{}3]$, $[t, x, y]$, $t = 0..1$, $\{[0, 0, 0]\}$, $x = 0..4$, $y = 0..4$,
scene $= [x, y]$, arrows = THIN, grid = $[20, 20]$);

because the trajectory through $[0, 0, 0]$ gives simply the single point $x = y = 0$ in the x, y phase plane. We have included the one initial point $[0, 0, 0]$ because the program calls for at least one. If the default grid is too coarse or too fine we can define it through the grid option, where grid = [20, 20] calls for a 20 by 20 grid.

Besides generating the x, y phase plane, phaseportrait can also generate plots of x or y versus t using the scene option. For instance,

phaseportrait($[y, -x + x\hat{}3]$, $[t, x, y]$, $t = 0..5$, $\{[0, 0.2, 0], [0, 0.8, 1.3]\}$,
stepsize $= 0.05$, scene $= [t, x]$);

gives two plots of $x(t)$ versus t over $0 \leq t \leq 5$, one with initial conditions $x(0) = 0.2$ and $y(0) = x'(0) = 0$, and the other with initial conditions $x(0) = 0.8$ and $y(0) = x'(0) = 1.3$.

EXERCISES 7.2

1. We stated, below (5) that if we solve (5) for y (i.e., dx/dt), separate variables, and integrate, we obtain the general solution $x(t) = A \sin(\omega t + \phi)$ of (1). Here we ask you to do that, to carry out those steps.

2. Supply the steps missing between the first and second lines of (18).

3. We found in Fig. 6 that for the hard-spring oscillator the frequency increases with the amplitude. Explain, in simple terms, why that result makes sense.

4. Determine the equation of the phase trajectories for the given system, and sketch several representative trajectories. Use arrows to indicate the direction of movement along those trajectories.

(a) $x' = y, \quad y' = -x$ (b) $x' = xy, \quad y' = -x^2$
(c) $x' = y^2, \quad y' = -xy$

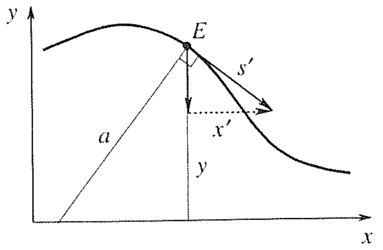
5. Determine the equation of the phase trajectories and sketch enough representative trajectories to show the essential features of the phase portrait. Use arrows to indicate the direction of movement along those trajectories.

(a) $x' = y, \quad y' = -y$ (b) $x' = y, \quad y' = y$
(c) $x' = y, \quad y' = x$ (d) $x' = y, \quad y' = 9x$
(e) $x' = x, \quad y' = x$ (f) $x' = x, \quad y' = -4x$

6. (*Period, for soft-spring oscillator*) In the paragraph below (21), we suggest that the period T of the periodic motions inside of $BDEHB$ (Fig. 8) tends to 2π in the limit as the amplitude A of the motion tends to zero, and to ∞ as $A \to 1$. Here we ask you to explore that claim with calculations. Specifically, use phaseportrait (or other software) to solve $x'' + x - x^3 = 0$ subject to the initial conditions $x(0) = A$, $x'(0) = 0$, for $A = 0.05, 0.3, 0.6, 0.9, 0.95, 0.99$. From your results, obtain the period T for each case and plot T versus A for those values of A (and additional ones if you wish). Does the claim made in the first sentence appear to be correct?

7. We stated, in our discussion of Fig. 8, that all trajectories outside of the "football" region correspond to nonperiodic motions. Explain why that is true.

8. (*Graphical determination of phase velocity*) (a) For the system (22), consider the special case where $P(x, y) = y$, as occurred in (3) and (7), for instance. From the accompanying sketch, show that in that case the phase velocity s' can be interpreted graphically as



$$s' = a, \tag{8.1}$$

where a is the perpendicular distance from E to the x axis. (b) Consider a rectangular phase trajectory $ABCDA$, where the corner points have the x, y coordinates $A = (-1, 1)$, $B = (3, 1)$, $C = (3, -1)$, $D = (-1, -1)$. Using (8.1), plot the graph of $x(t)$ versus t, from $t = 0$ through $t = 20$, if the representative point E is at A at $t = 0$.

(c) Consider a phase trajectory ABC consisting of straight-line segments from $A = (-1, 0)$ to $B = (0, 1)$ to $C = (1, 0)$ with E at B at $t = 0$. Using (8.1), sketch the graph of $x(t)$ versus t over $-\infty < t < \infty$. Also, give $x(t)$ analytically over $-\infty < t < 0$ and $0 < t < \infty$.
(d) Consider a straight-line phase trajectory from $A = (0, 5)$ to $B = (10, -5)$. Using (8.1), sketch the graph of $x(t)$ versus t over $0 < t < \infty$, if E is at A at $t = 0$.
(e) Same as (d), but with E at B at $t = 0$.

9. (a) Reduce the equation $x'' + 2x^3 = 0$ to a system of equations by setting $x' = y$. Find the equation of the phase trajectories and sketch several of them by hand. Show that for larger and larger motions the trajectories are flatter and flatter over $-1 < x < 1$.
(b) Use the *Maple* phaseportrait command (or other software) to generate the phase portrait and, on the same plot, the lineal element field, using barbed arrows to show the flow direction. You will need to make decisions, with some experimentation, as to the t-interval, the step size, the initial points, and so on, so as to obtain good results.

10. Reduce the equation $x'' + x^2 = 0$ to a system of equations by setting $x' = y$. Find the equation of the trajectories and carefully sketch seven or eight of them, so as to show clearly the key features of the phase portrait. Pay special attention to the one through the origin, and give its equation.

11. (*Volterra problem*) Consider the Volterra problem (24), with $\alpha = \beta = \gamma = \delta = 1$. Determine any fixed points. Use phaseportrait (or other software) to obtain the lineal element field, with barbed arrows, over the region $0 < x < 4$ and $0 < y < 4$, say. (Of course, x and y need to be positive because they are populations.). On that plot, sketch a number of representative trajectories. You should find a circulatory motion about the point $(1, 1)$. Can you tell, from the lineal element field, whether the trajectories circulate in closed orbits or whether they spiral in (or away from) that point?

7.3 Singular Points and Stability

7.3.1. Existence and uniqueness. Before we go any further, we need to return to the question of the existence and uniqueness of solutions. Theorem 2.4.1 gives conditions on f for the equation $y' = f(x, y)$ to admit a unique solution through a given initial point $y(a) = b$. Theorem 3.9.1 does likewise for a system of such equations, but covered only linear equations. In the present chapter, however, our principal interest is in systems of nonlinear equations so we give the following result, which is essentially a generalization of Theorem 2.4.1 to a system of equations.

THEOREM 7.3.1 *Existence and Uniqueness*
Let $f(x, y, t)$ and $g(x, y, t)$ and each of the partial derivatives f_x, f_y, g_x, g_y exist and be continuous in some neighborhood of the point (x_0, y_0, t_0) in Cartesian x, y, t space. Then the initial-value problem

$$\frac{dx}{dt} = f(x, y, t); \qquad x(t_0) = x_0 \tag{1a}$$

$$\frac{dy}{dt} = g(x, y, t); \qquad y(t_0) = y_0 \tag{1b}$$

has a solution $x(t)$, $y(t)$ on some open t interval containing $t = t_0$ and that solution is unique.

More general and more powerful theorems could be given but this one will suffice for our purposes in this chapter. For such theorems and proofs we refer you to texts on differential equations such as G. Birkhoff and G.-C. Rota, *Ordinary Differential Equations*, 2nd ed. (New York: John Wiley, 1969).

For instance, consider the soft-spring oscillator equation $x'' + x - x^3 = 0$ or, equivalently, the system

$$x' = y, \tag{2a}$$
$$y' = -x + x^3 \tag{2b}$$

that we studied in Section 7.2. In this case $f(x, y, t) = y$, $g(x, y, t) = -x + x^3$, $f_x = 0$, $f_y = 1$, $g_x = -1 + 3x^2$, $g_y = 0$ are continuous for all values of x, y, and t, so Theorem 7.3.1 assures us that no matter what initial condition is chosen there is a unique solution through it. The extent of the t interval over which that solution exists is not predicted by the theorem, which is a "local" theorem like Theorem 2.4.1. But it is understood that that interval is not merely the point t_0 itself, for how could dx/dt and dy/dt make sense if $x(t)$ and $y(t)$ were defined only at a single point? Linear differential equations are simpler, and for them we have "global" theorems such as Theorem 3.9.1.

If f and g satisfy the conditions of Theorem 7.3.1 at (x_0, y_0, t_0) in x, y, t space, then there does *exist* a solution curve, or trajectory, through that point, and there is only *one* such trajectory. Geometrically, it follows that trajectories in x, y, t space cannot touch or cross each other at a point of existence and uniqueness. However, what about the possibility of crossings of trajectories in the x, y phase plane? Be careful, because whereas the theorem precludes crossings in three-dimensional x, y, t space, the phase plane shows only the projection of the three-dimensional trajectories onto the two-dimensional x, y plane. For instance, choose any point P_0 on a closed orbit inside the "football" in Fig. 8 of Section 7.2. As the representative point P goes round and round on that orbit it passes through P_0 an infinite number of times, yet that situation does not violate the theorem because if that trajectory is viewed in three-dimensional x, y, t space, we see that it is actually helical, and there are no self-crossings. The only points of serious concern in Fig. 8 are $(1, 0)$ and $(-1, 0)$. But here too there is no violation of the theorem because there is only the unique trajectory $x(t) = 1$ and $y(t) = 0$ through any initial point $(1, 0, t_0)$ – namely, a straight-line trajectory which is perpendicular to the x, y plane and which extends from $-\infty$ to $+\infty$ in the t direction. The trajectories DE and IE, in the x, y, t space, approach that line asymptotically as $t \to \infty$, and the trajectories FE and HE approach it asymptotically (both in x, y, t space and in the x, y phase plane) as $t \to -\infty$, but they never reach it. Similarly for $(-1, 0, t_0)$.

Recall, from Section 7.2, that we proceeded to divide (2b) by (2a), obtaining

$$\frac{dy}{dx} = \frac{-x + x^3}{y}. \tag{3}$$

Thus, apart from the question of existence and uniqueness for the system (2) in x, y, t space, there is the separate question as to the existence and uniqueness of the solution of the single equation (3) through a given initial point in the x, y plane, to which question we now turn. Here, $f(x,y) = (-x+x^3)/y$ and $f_y = (x-x^3)/y^2$ are continuous everywhere except on the line $y = 0$, that is, on the x-axis. However, since we are not aiming at finding $y(x)$ but at finding the phase trajectories, then it doesn't matter if, at any given location, we regard y as a function of x or vice versa. Thus, if we re-express (3) as $dx/dy = y/(-x+x^3)$, then we find that the right side and its partial derivative with respect to the dependent variable x are continuous everywhere except at $x = 0, 1$, and -1. Thus, the only points in the phase plane at which there can possibly be breakdowns of existence and uniqueness of a trajectory passing through that point are the intersections of the horizontal line $y = 0$ and the vertical lines $x = 0, 1$, and -1, namely, the points $(-1,0), (0,0)$, and $(1,0)$, and we call these "singular points" of (3). At each of these points dy/dx and dx/dy are of the form $0/0$, so there is no unique slope defined through that point. Indeed we do find breakdowns of uniqueness at the points $(-1,0)$ and $(1,0)$ in Fig. 8 because at each point we find two crossing trajectories. (Remember that we are in x, y space now, not x, y, t space, and that time does not enter explicitly.) And at $(0,0)$ we find a failure of existence because there is no solution that passes through that point.

The upshot is that even though the system (2) has a unique solution in x, y, t space, through any point $(-1,0,t_0)$, $(0,0,t_0)$, or $(1,0,t_0)$, the equation (3) does *not* have a unique solution through any of the points $(-1,0), (0,0)$, and $(1,0)$. We therefore say that these three points are singular points of (3), which concept we now formalize.

7.3.2. Singular points. Thus, in considering the general case

$$\frac{dy}{dx} = \frac{f(x,y)}{g(x,y)}, \tag{4}$$

obtained from $x' = dx/dt = f(x,y)$ and $y' = dy/dt = g(x,y)$ by division, we say that any point x_s, y_s at which both $f(x,y)$ and $g(x,y)$ vanish is a **singular point** (or **critical point**) of (4). Recall that from a dynamical point of view a singular point is an equilibrium point, or fixed point, because if we start at such a point then we remain there for all t because $x' = y' = 0$ there.

Singular points are of great importance in phase plane analysis. They need not be isolated. For example, if $f(x,y) = x$ and $g(x,y) = x$, then every point on the y axis (i.e., the line $x = 0$) is a singular point. To study a singular point of a nonlinear system, we focus attention on its immediate neighborhood by expanding $f(x,y)$ and $g(x,y)$ in Taylor series about that point and linearizing, that is, cutting them off after the first-order terms. For instance, to study the singular point of (2) at $(1,0)$, write

$$x' = y, \tag{5a}$$

$$y' = -x + x^3 = 0 + 2(x-1) + \frac{6}{2!}(x-1)^2 + \frac{6}{3!}(x-1)^3. \tag{5b}$$

Close to $(1, 0)$ we neglect the higher-order terms and consider the linearized version

$$x' = y, \tag{6a}$$
$$y' = 2(x-1) \tag{6b}$$

or, moving our coordinate system to $(1, 0)$ for convenience by setting $X = x - 1$ and $Y = y$,

$$X' = Y, \tag{7a}$$
$$Y' = 2X. \tag{7b}$$

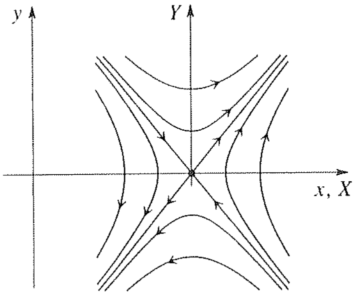


Figure 1. The flow near the singular point $(1, 0)$.

Dividing these gives $dY/dX = 2X/Y$, with the solution $Y^2 = 2X^2 + C$; $C = 0$ gives $Y = \pm\sqrt{2}X$; and for $C \neq 0$ we have $Y = \pm\sqrt{(2X^2 + C)}$. These curves are shown in Fig. 1. The pairs crossing the X axis correspond to increasingly negative values of C, and those crossing the Y axis correspond to increasingly positive values of C.

Similarly, to study the singular point $(0, 0)$ observe that the right-hand sides of (2a) and (2b) are already Taylor series expansions about $x = 0$ and $y = 0$. Thus, keeping terms only up to first order gives the linearized version

$$x' = y, \tag{8a}$$
$$y' = -x, \tag{8b}$$

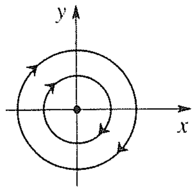


Figure 2. The flow near the singular point (0,0).

with $dy/dx = -x/y$ giving the family of circles $y^2 + x^2 = C$ and hence the trajectories shown in Fig. 2.

Treating the singular point $(-1, 0)$ in the same manner, we find the same behavior there as at $(1, 0)$. If we show the three results together, in Fig. 3, it is striking how those three localized phenomena appear to set up the global flow in the whole plane. That is, if we fill in what's missing, by hand, we obtain – at least in a qualitative or topological sense – the same picture as in Fig. 8 of Section 7.2.*

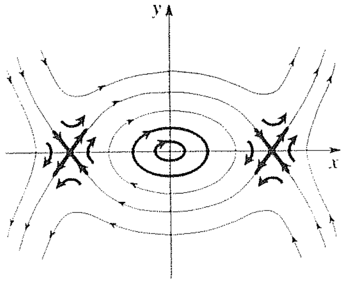


Figure 3. Global flow determined, qualitatively, by the singular points.

From this example we can see some things and raise some questions. We see that by virtue of our Taylor expansions of $f(x, y)$ and $g(x, y)$ about the singular point (x_s, y_s) and their linearization, we are always going to end up with linearized equations of the form

$$X' = aX + bY, \tag{9a}$$
$$Y' = cX + dY \tag{9b}$$

to study, where $X \equiv x - x_s$ and $Y \equiv y - y_s$ are Cartesian coordinate axes located at the singular point. Thus, we might as well study the general system (9) once and for all. Evidently, for different combinations of a, b, c, d there can be different types of singular points, for from Fig. 3 it seems clear that the ones at $(1, 0)$ and $(-1, 0)$ are different from the one at $(0, 0)$. How many different types are there?

*It may appear inconsistent that the trajectories near the origin in Fig. 3 look elliptical, whereas they are circles in Fig. 2. That distortion, from circles to ellipses, is merely the result of stretching the x axis relative to y axis for display purposes.

What are they?

7.3.3. The elementary singularities and their stability. We wish to solve (9), examine the results, and classify them into types. If we equate the right-hand sides of (9) to zero, we have the unique solution $X = Y = 0$ only if the determinant $ad - bc$ is nonzero. If that determinant vanishes, then the solution $X = Y = 0$ is nonunique, and there is either an entire line of solutions through the origin or the entire plane of solutions. For instance, if $a = b = 0$ and c and d are not both zero, then every point on the line $cX + dY = 0$ is a singular point of (9), and if $a = b = c = d = 0$, then every point in the plane is a singular point of (9). Wishing to study the generic case, where the origin is an *isolated* singular point, we will require of a, b, c, d that $\mathbf{ad - bc \neq 0}$.

If we solve (9), for instance by elimination, we find that the solution is of the form

$$X(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}, \quad Y(t) = C_3 e^{\lambda_1 t} + C_4 e^{\lambda_2 t}, \tag{10a,b}$$

where C_1, C_2, C_3, C_4 are not independent, and where λ_1, λ_2 are the roots

$$\lambda = \frac{a + d \pm \sqrt{(a-d)^2 + 4bc}}{2} \tag{11}$$

of the characteristic equation

$$\lambda^2 - (a+d)\lambda + (ad - bc) = 0. \tag{12}$$

Since $ad - bc \neq 0$, zero is not among the roots. There are exactly four possibilities:

(1) purely imaginary roots (CENTER),
(2) complex conjugate roots (FOCUS),
(3) real roots of the same sign (NODE),
(4) real roots of opposite sign (SADDLE).

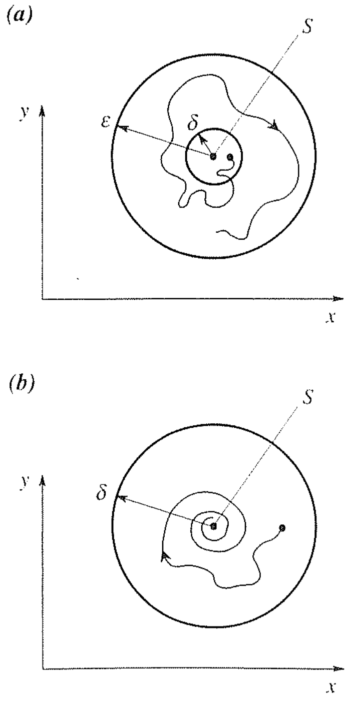


Figure 4. Stability and asymptotic stability.

These cases lead to four different types of singularity: center, focus, node, and saddle, as we note within parentheses, and we will discuss these in turn. In doing so, it is important to examine each in terms of its stability, which concept we define before continuing.

A singular point $S = (x_s, y_s)$ of the autonomous system (4) is said to be **stable** if motions (i.e., trajectories) that start out sufficiently close to S remain close to S. To make that intuitively stated definition mathematically precise, let $d(P_1, P_2)$ denote the distance* between any two points $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. Further, we continue to let $P(t) = (x(t), y(t))$ denote the representative point in the phase plane corresponding to (4). Then, a singular point S is stable if, given any $\epsilon > 0$ (i.e., as small as we wish) there is a $\delta > 0$ such that $d(P(t), S) < \epsilon$ for all $t > 0$ if $d(P(0), S) < \delta$. (See Fig. 4a.) If S is not stable, then it is **unstable**.

*In the Euclidean sense, the distance $d(P_1, P_2)$ is defined as $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$, but one can define distance in other ways. Here, we understand it in the Euclidean sense.

Further, we say that S is not only stable but **asymptotically stable** if motions that start out sufficiently close to S not only stay close to S but actually approach S as $t \to \infty$. That is, if there is a $\delta > 0$ such that $d(P(t), S) \to 0$ as $t \to \infty$ whenever $d(P(0), S) < \delta$, then S is asymptotically stable. (See Fig. 4b.)

Now let us return to the four cases listed. The most inciteful way to study these cases is by seeking solutions in exponential form and dealing with the "eigenvalue problem" that results. However, the eigenvalue problem is not discussed until Chapter 11, so in the present section we rely on an approach that should suffice but which is in some ways less satisfactory. In Section 11.5 we return to this problem and deal with it as an eigenvalue problem. If you are already sufficiently familiar with linear algebra, we suggest that you study that section immediately following this one.

It is convenient to use the physical example of the mechanical oscillator, with the governing equation $mx'' + px' + kx = 0$, or

$$x' = y, \tag{13a}$$

$$y' = -\frac{k}{m}x - \frac{p}{m}y \tag{13b}$$

as a unifying example because by suitable choice of m, p, k we can obtain each of the four cases, and because this application has already been discussed in Section 3.5. [Here we use p instead of c for the damping coefficient to avoid confusion with the c in (9b).]

Purely imaginary roots. (CENTER) Let $p = 0$ so there is no damping. Then $a = 0$, $b = 1$, $c = -k/m$, $d = 0$; (11) gives the purely imaginary roots $\lambda = \pm i\sqrt{k/m}$ and $dy/dx = -(k/m)x/y$ gives the family of ellipses

$$\frac{1}{2}my^2 + \frac{1}{2}kx^2 = C \tag{14}$$

sketched in Fig. 5a. The singular point at $(0, 0)$ is called a **center** because it is surrounded by closed orbits corresponding to periodic motions. For instance, with $\lambda_1 = +i\omega$ (where $\omega = \sqrt{k/m}$ is the natural frequency) and $\lambda_2 = -i\omega$, (10a) gives $x(t) = C_1 \exp(i\omega t) + C_2 \exp(i\omega t)$ or, equivalently,

$$x(t) = A \sin(\omega t + \phi). \tag{15}$$

(Here, $X = x$ and $Y = y$ because the singular point is at $x = y = 0$.) In Fig. 5a the principal axes of the elliptical orbits coincide with the x, y coordinate axes. More generally, they need not. For instance, for the system

$$x' = \frac{\sqrt{8}}{3}x + \frac{4}{3}y, \tag{16a}$$

$$y' = -\frac{11}{3}x - \frac{\sqrt{8}}{3}y \tag{16b}$$

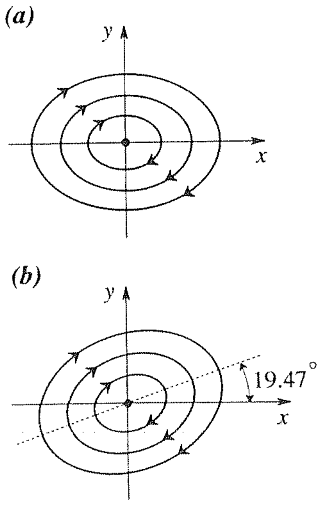


Figure 5. A center at $(0, 0)$.

(11) again gives purely imaginary roots, $\lambda = \pm i/\sqrt{3}$ so the solutions are harmonic oscillations with frequency $1/\sqrt{3}$, but the principal axes of the elliptical orbits are at an angle of $\sin^{-1}(1/3) = 19.47^\circ$ with respect to the x, y axes as shown in Fig. 5b (see Exercise 5). [The system (16) is, of course, not a special case of (13), it is a separate example.]

We see that a center is stable but not asymptotically stable.

Complex conjugate roots. (FOCUS) This time let p be positive in (13), but small enough so that $p < \sqrt{4km}$. According to the terminology introduced in Section 3.8, we say that the damping is subcritical because $p_{cr} = \sqrt{4km}$. Then $a = 0$, $b = 1$, $c = -k/m$, $d = -p/m$; (11) gives the complex conjugate roots

$$\lambda = -\frac{p}{2m} \pm \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}} = -\frac{p}{2m} \pm i\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2},$$

and (10a) gives the solution

$$\begin{aligned} x(t) &= e^{-pt/2m}\left[A\cos\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\right)t + B\sin\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\right)t\right] \\ &= Ce^{-pt/2m}\sin\left(\sqrt{\frac{k}{m} - \left(\frac{p}{2m}\right)^2}\,t + \phi\right), \end{aligned} \tag{17}$$

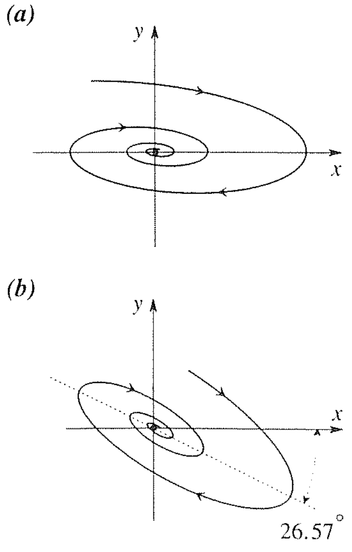


Figure 6. A stable focus at (0, 0).

where C and ϕ are arbitrary constants. As we discussed in Section 3.8, this solution differs from the undamped version (15) in two ways. First, the frequency of the sinusoid is diminished from the natural frequency $\omega = \sqrt{k/m}$ to $\sqrt{k/m - (p/2m)^2}$, and the $\exp(-pt/2m)$ factor modulates the amplitude, reducing it to zero as $t \to \infty$. In terms of the phase portrait, one obtains a family of spirals such as the one shown in Fig. 6a. If we imagine the representative point P moving along that curve, we see that the projection onto the x axis is indeed a damped oscillation. We call the singularity at the origin a **focus** because trajectories "focus" to the origin as $t \to \infty$; the term **spiral** is also used.

In Fig. 6a the principal axes of the orbits, which would be elliptical if not for the damping, coincide with the x, y coordinate axes. More generally, they need not. For instance, for the system

$$x' = \frac{1}{3}x + \frac{7}{9}y, \tag{18a}$$

$$y' = -\frac{1}{3}x - \frac{4}{9}y \tag{18b}$$

we obtain similar results but with the principal axes rotated clockwise by an angle of $\sin^{-1}(1/\sqrt{5}) = 26.57^\circ$ as shown in Fig. 6b.

In each case (Fig. 6a and 6b) we see that the focus is stable and, indeed, asymptotically stable as well. However, one can have foci that wind outward instead of inward, and these will be unstable. For instance, if we return to the solution (17) of the damped oscillator system (13), but this time imagine p to be negative (without concerning ourselves with how that might be arranged physically), and smaller in magnitude than $\sqrt{4km}$ as before, then in place of the clockwise inward flow shown in Fig. 5a, we obtain the counterclockwise outward flow shown in Fig. 7, and we classify the singularity at the origin as an unstable focus. Note that a stable singular point can, additionally, be asymptotically stable or not, but an unstable singular point is simply unstable.

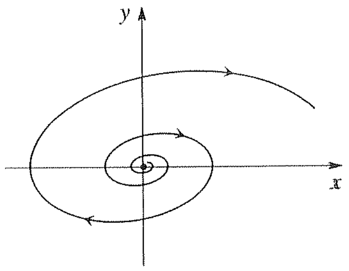


Figure 7. An unstable focus at$(0,0)$.

Real roots of the same sign. (NODE) We've seen that without damping the mechanical oscillator (13) gives pure oscillations, elliptical orbits, and a center. With light damping (i.e., $0 < p < p_{cr}$) it gives damped oscillations and a stable focus. If we now increase p so as to exceed p_{cr}, then the oscillations disappear altogether, and we have the solution form

$$x(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} \tag{19a}$$

with

$$\lambda_1 = -\frac{p}{2m} + \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}}, \qquad \lambda_2 = -\frac{p}{2m} - \sqrt{\left(\frac{p}{2m}\right)^2 - \frac{k}{m}}.$$

Because of the way we have numbered the λ's, we have $\lambda_2 < \lambda_1 < 0$. Since $x' = y$ in this application we have from (19a),

$$y(t) = \lambda_1 C_1 e^{\lambda_1 t} + \lambda_2 C_2 e^{\lambda_2 t}. \tag{19b}$$

We can see from (19) that

$$x(t) \sim C_1 e^{\lambda_1 t}, \qquad y(t) \sim \lambda_1 C_1 e^{\lambda_1 t} \tag{20}$$

and

$$y \sim \lambda_1 x \tag{21}$$

as $t \to \infty$, provided that the initial conditions do not give $C_1 = 0$. If they do give $C_1 = 0$, then

$$x(t) = C_2 e^{\lambda_2 t}, \qquad y(t) = \lambda_2 C_2 e^{\lambda_2 t} \tag{22}$$

and

$$y = \lambda_2 x \tag{23}$$

as $t \to \infty$.

The resulting phase portrait is shown in Fig. 8a. We call the singularity at $(0,0)$ a **node** – more specifically, an **improper node** ("improper" is explained below). In accord with the preceding discussion, observe from Fig. 8a that in the exceptional case in which the initial point lies on the line of slope λ_2 the representative point approaches the origin along that line. In all other cases, the approach is asymptotic to the line of slope λ_1. If we let p tend to p_{cr}, then the two lines coalesce

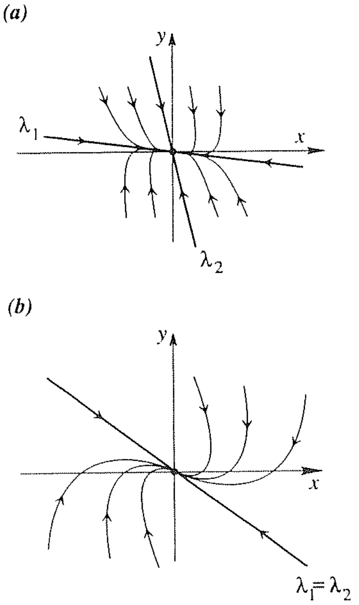


Figure 8. A stable improper node at $(0,0)$: distinct roots and repeated roots, respectively.

and we obtain the portrait shown in Fig. 8b, which is, likewise, an improper node (Exercise 6).

If p is negative and greater in magnitude than p_{cr}, then we see from the expressions given above for the λ's, that $\lambda_1 > \lambda_2 > 0$, and the phase portrait is as shown in Fig. 9. The nodes shown in Fig. 8a,b are both stable, asymptotically stable, and the one shown in Fig. 9 (which is analogous to the one in Fig. 8a) is unstable.

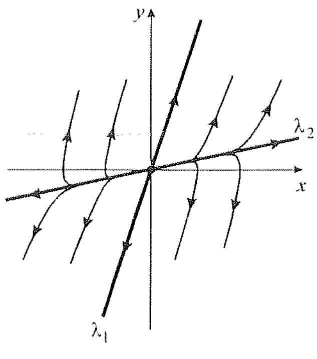


Figure 9. An unstable improper node at $(0, 0)$: distinct roots.

Our mechanical oscillator example is but one example leading to a node. We could make up additional ones by writing down equations $x' = ax + by$ and $y' = cx + dy$ if we choose the coefficients a, b, c, d so that the two λ's are of the same sign, but the results will be one of the types shown in Fig. 8 and 9. There is, however, another type of node, which we illustrate by the problem

$$x' = ax, \tag{24a}$$

$$y' = ay. \tag{24b}$$

That is, $b = c = 0$ and $a = d$. Then $\lambda_1 = \lambda_2 = a$, and we have the solution

$$x(t) = Ae^{at}, \tag{25a}$$

$$y(t) = Be^{at}, \tag{25b}$$

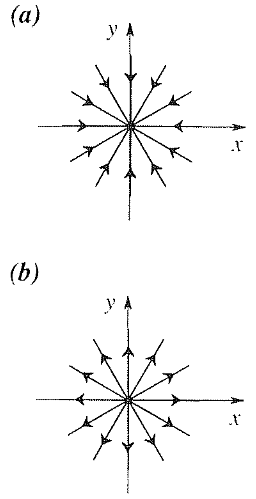


Figure 10. Stable and unstable proper nodes at $(0, 0)$.

where A, B are arbitrary. In this case y/x is not only asymptotic to a constant as $t \to \infty$, it is *equal* to a constant for all t. Thus, the phase portrait is as shown in Fig. 10a if $a < 0$, and as in Fig. 10b if $a > 0$. The former is an asymptotically stable node, and the latter is an unstable node. But this time we call them **proper nodes** (or **stars**) because every trajectory is a straight line through the singular point $(0, 0)$, not just one or two of them.

Real roots of opposite sign. (SADDLE) Consider, once again, the undamped mass/spring system governed by the equation $mx'' + kx = 0$, or

$$x' = y, \tag{26a}$$

$$y' = -\frac{k}{m}x. \tag{26b}$$

This time, imagine k to be negative (without concern about how that case could occur physically) and set $-k/m \equiv h^2$. Then (26) gives $dy/dx = h^2x/y$ so

$$y^2 = h^2x^2 + C \qquad \text{or} \qquad y(x) = \pm\sqrt{h^2x^2 + C}, \tag{27}$$

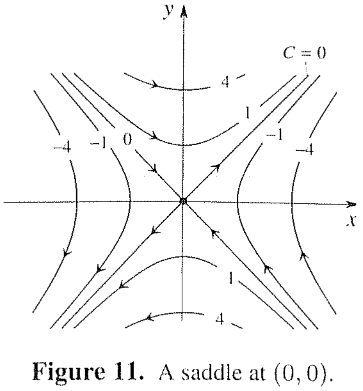


Figure 11. A saddle at $(0, 0)$.

which trajectories are shown in Fig. 11 for various values of C. In particular, $C = 0$ gives the two straight lines through the singular point, namely, $y = \pm hx$, with the flow approaching the origin on one ($y = -hx$) and moving away from it on the other ($y = +hx$). Such a singular point is called a **saddle** and is always unstable. The two straight-line trajectories through the saddle, along which the flow is attracted and repelled, are called the **stable** and **unstable manifolds**, respectively.

Of course, (26) is not the only example of a linear system $x' = ax + by$ and $y' = cx + dy$ with a saddle. Any such system with real roots of opposite sign will have such a singularity. For instance,

$$x' = x + 2y, \tag{28a}$$
$$y' = 8x - 5y \tag{28b}$$

has the roots $\lambda = 3$ and $\lambda = -7$. Thus, it has a saddle, and we know that two straight-line solutions can be found through the origin. To find them, try $y = \kappa x$. Putting $y = \kappa x$ into (28) gives $x' = (1 + 2\kappa)x$ and $x' = (8 - 5\kappa)x/\kappa$ so it follows from these that we need $1 + 2\kappa = (8 - 5\kappa)/\kappa$, which equation gives the slopes $\kappa = 1$ and $\kappa = -4$. (If we would obtain $\kappa = \infty$, we would understand that, from $y = \kappa x$, to correspond to the x axis.) With $\kappa = 1$ the equation $x' = (1+2\kappa)x = 3x$ gives $x(t)$ proportional to $\exp(3t)$ [likewise for $y(t)$ because $y = \kappa x$], and with $\kappa = -4$ it gives $x(t)$ proportional to $\exp(-7t)$. Thus, the line trajectory $y = x$ is the unstable manifold (since x and y grow exponentially on it), and the line trajectory $y = -4x$ is the stable manifold (since x and y die out exponentially on it).

The same procedure, which we have just outlined and which should be clearly understood, can be used for a node as well, to find any straight-line trajectories through the node.

7.3.4. Nonelementary singularities. In this final subsection we turn from the elementary singularities to nonelementary ones, with two purposes in mind. First, one doesn't completely understand elementary singularities until one distinguishes elementary singularities from nonelementary ones and, second, nonelementary singularities do arise in applications.

Recall that

$$X' = aX + bY, \quad Y' = cX + dY \tag{29a,b}$$

has an elementary singularity at $(0, 0)$ if $ad - bc \neq 0$. Consider two examples.

EXAMPLE 1. The system

$$x' = y, \quad y' = y \tag{30a,b}$$

has the phase trajectories $y = x + C$ and the phase portrait shown in Fig. 12. Since $ad - bc = (0)(1) - (1)(0) = 0$, the singularity of (30) at $(0, 0)$ is nonelementary. It is nonisolated and, in fact, $y = 0$ is an entire line of singular points. ∎

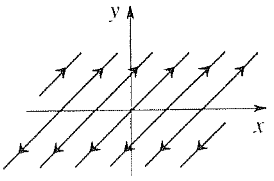


Figure 12. Nonelementary singularity of (30) at $(0, 0)$.

EXAMPLE 2. Consider the singularity of the system

$$x' = y, \quad y' = 1 - \cos x \tag{31a,b}$$

at $(0, 0)$. Expanding the right side of (31b) gives $1 - \cos x = \frac{1}{2}x^2 - \frac{1}{24}x^4 + \cdots$ so the linearized version of (31) is $x' = 0x + 1y$ and $y' = 0x + 0y$. Thus, $ad - bc = (0)(0) - (1)(0) = 0$ again and the singularity of (31) at the origin is nonelementary. The difficulty this time is not that the singular point is not isolated; it *is*. The problem is that it is of higher order for when we linearize the expansion of $1 - \cos x$ we simply have $0x + 0y$. In not retaining at least the first nonvanishing term (namely, $x^2/2$) we have "thrown out the baby with the bathwater." To capture the local behavior of (31) near the origin, we need to retain that leading term and consider the system

$$x' = y, \quad y' = \frac{1}{2}x^2. \tag{32a,b}$$

Dividing (32b) by (32a) and integrating gives $y = \sqrt{\dfrac{x^3}{3} + C}$, several of which trajectories are shown in Fig. 13. We see that the singularity is, indeed, isolated, but that the phase portrait is not of one of the elementary types. ▮

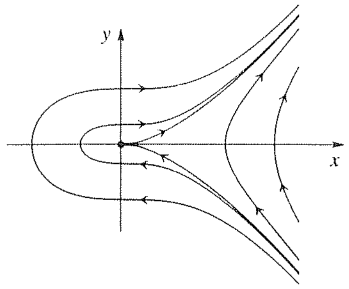


Figure 13. Nonelementary singularity of (31) at (0, 0).

Closure. In this section we establish a foundation for our use of the phase plane in studying nonlinear systems. We begin with the issue of existence and uniqueness of solutions, first in x, y, t space (Theorem 7.3.1), and then in the x, y phase plane. The latter leads us to introduce the concept of a singular point in the phase plane as a point at which both $x' = P(x, y) = 0$ and $y' = Q(x, y) = 0$. To study a singular point $S = (x_s, y_s)$ one focuses on the immediate neighborhood of S, in which neighborhood we work with the locally linearized equations $X' = aX + bY$ and $Y' = cX + dY$, where $X = x - x_s$ and $Y = y - y_s$, so X, Y is a Cartesian system with its origin at S. Studying that linearized system, we categorize the possible "flows" into four qualitatively distinct types – the center, focus, node, and saddle – and illustrate each through the mass/spring system $mx'' + px' + kx = 0$, with suitable choices of m, p, k, and other examples as well. These are the so-called elementary singularities that result when $ad - bc \neq 0$. In the next section we apply these results to several nonlinear systems, where we will see the role of such singular points in establishing the key features of the overall flow in the phase plane.

EXERCISES 7.3

1. Find all singular points of the given system. Are they isolated?

(a) $x' = 0$
$y' = 2x - y$

(b) $x' = 2x - 2y$
$y' = x - y$

(c) $x' = x^2 + y^2 - 1$
$y' = x - y$

(d) $x' = \sin y$
$y' = x + y$

(e) $x' = x^2 + y^2$
$y' = x + y$

(f) $x' = 1 - e^y$
$y' = 1 - x^2 - x \sin y$

(g) $x' = xy - 4$
$y' = x - 2y$

(h) $x' = \cos(x - y)$
$y' = xy - 1$

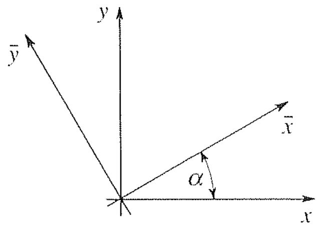
2. Is it possible for (4) to have *no* singular points? Explain.

3. Derive the solution (10) of (9) by the method of elimination, and find C_3 and C_4 in terms of C_1 and C_2.

4. Suppose that we reverse the ϵ's and δ's in our definition of stability, so that the definition becomes: A singular point S is stable if, given any $\epsilon > 0$ (i.e., as small as we wish), there is a $\delta > 0$ such that $d(P(t), S) < \delta$ for all $t > 0$ if $d(P(0), S) < \epsilon$. Would that definition work? That is, would it satisfy the idea of motions that start out sufficiently close to S remaining close to S? Explain.

5. In this exercise we wish to elaborate on our claim below (16) that the principal axes of the elliptical orbits are at an angle of 19.47° with respect to the x, y axes as shown in Fig. 5b.

(a) If x, y and $\overline{x}$, $\overline{y}$ coordinate systems are at an angle α, as shown here, show that the x, y and $\overline{x}$, $\overline{y}$ coordinates of any given point are related according to



$$x = \overline{x}\cos\alpha - \overline{y}\sin\alpha, \tag{5.1a}$$

$$y = \overline{x}\sin\alpha + \overline{y}\cos\alpha \tag{5.1b}$$

(b) Putting (5.1) into (16), insist upon the result being of the form

$$\overline{x}' = \beta^2\overline{y}', \qquad \overline{y}' = -\gamma^2\overline{x}' \tag{5.2}$$

for some constants β and γ so as to yield elliptical orbits with $\overline{x}, \overline{y}$ as principal axes, and show that you obtain $\alpha = 19.47°$. If you obtain another α as well, explain its significance.

6. We claimed that if $p = p_{cr}$, then the two straight-line trajectories in Fig. 8a coalesce, as shown in Fig. 8b. Here, we ask you to verify that claim. Begin by recalling that if $\lambda_1 = \lambda_2 = \lambda$, then the general solution of (13) is

$$x(t) = (C_1 + C_2 t)e^{\lambda t}, \qquad y(t) = x'(t) = \text{etc.}$$

7. What does Fig. 8a look like in the limit as $p \to \infty$? Sketch it.

8. Given the presence of the saddles (i.e., saddle-type singularities) at $(-1, 0)$ and $(1, 0)$ and the center at $(0, 0)$, can you come up with any global flow patterns that are qualitatively different from the one sketched in Fig. 3? Explain. (Assume that these three are the only singularities.)

9. (*Saddles and nodes*) Classify the singularity at the origin, find the equations of any straight-line trajectories through the origin, and sketch the phase portrait, including flow direction arrows.

(a) $x' = x + y$, $y' = 4x + y$
(b) $x' = y$, $y' = -x - 4y$
(c) $x' = x + 2y$, $y' = x - 2y$
(d) $x' = -x + 3y$, $y' = x - y$
(e) $x' = 3x + y$, $y' = -x + y$
(f) $x' = -3x + y$, $y' = -x - y$
(g) $x' = 2x + y$, $y' = x + 2y$
(h) $x' = x + 3y$, $y' = 3x + y$
(i) $x' = x + y$, $y' = x + 2y$
(j) $x' = -3x + y$, $y' = x - 3y$

10. Prove that a linear system $x' = ax + by$, $y' = cx + dy$ can have one, two, or an infinite number of straight-line trajectories through the origin, but never a finite number greater than two.

11. Classify the singularity at the origin as a center, focus, node, or saddle. If it is a focus, node, or saddle, then classify it, further, as stable or unstable.

(a) $x' = x - 4y$, $y' = x + y$
(b) $x' = 2x + 3y$, $y' = 2x - y$
(c) $x' = x + y$, $y' = x - 4y$
(d) $x' = x + 3y$, $y' = -x - y$
(e) $x' = -2x - 3y$, $y' = 3x + 2y$
(f) $x' = -x + y$, $y' = -x - 2y$
(g) $x' = 2x - y$, $y' = -x + 3y$
(h) $x' = -2x - y$, $y' = -x - 3y$

12. (a)–(h) Use computer software to obtain the phase portrait for the corresponding system in Exercise 11. Be sure to include any key trajectories – namely, any straight-line trajectories through the origin. From the phase portrait, classify the singularity as a center, focus, node, or saddle, state whether it is stable, asymptotically stable, or unstable, and use arrows to show the flow direction.

7.4 Applications

In Sections 7.2 and 7.3 we established the phase plane concept, the idea of singularities, the center, focus, node, and saddle singularities of linear systems, and their stability or instability. With those fundamentals in hand, we can undertake some interesting applications.

7.4.1. Singularities of nonlinear systems. We are interested in the autonomous system

$$x' = P(x, y), \tag{1a}$$
$$y' = Q(x, y). \tag{1b}$$

The x, y points at which both $P(x, y) = 0$ and $Q(x, y) = 0$ are the singular points of (1). Suppose we have determined those singular points, if any. We have emphasized the importance of determining the local flow near each such point. To do that, it should suffice to expand P and Q in Taylor series about the given singular point, say $S = (x_s, y_s)$, and to retain terms only up to the first order. Thus, instead of studying the complete equations (1) near S, we intend to simplify them by linearizing them about S.

Though familiar with the Taylor series of a function of one variable, from the calculus, you may not be familiar with Taylor series for functions of two or more variables. The Taylor series expansion of a given function of *two* variables, $f(x, y)$, about any point (a, b), is

$$\begin{aligned} f(x, y) = f(a, b) &+ \frac{1}{1!}\left[f_x(a, b)(x - a) + f_y(a, b)(y - b)\right] \\ &+ \frac{1}{2!}\left[f_{xx}(a, b)(x - a)^2 + 2f_{xy}(a, b)(x - a)(y - b) + f_{yy}(a, b)(y - b)^2\right] \\ &+ \text{terms of third order and higher,} \end{aligned} \tag{2}$$

where $(x - a)^m(y - b)^n$ is said to be of order $m + n$. One way to derive (2) is to expand $f(x, y)$ in one variable at a time. Thus, holding y fixed for the moment, and expanding in x about $x = a$, we know from the Taylor series formula for a function of a single variable that

$$f(x, y) = f(a, y) + f_x(a, y)(x - a) + \frac{1}{2!}f_{xx}(a, y)(x - a)^2 + \cdots. \tag{3}$$

The coefficients are functions of y alone, and each can now be expanded in y about $y = b$:

$$\begin{aligned} f(a, y) &= f(a, b) + f_y(a, b)(y - b) + \frac{1}{2!}f_{yy}(a, b)(y - b)^2 + \cdots, \\ f_x(a, y) &= f_x(a, b) + f_{xy}(a, b)(y - b) + \frac{1}{2!}f_{xyy}(a, b)(y - b)^2 + \cdots, \end{aligned} \tag{4}$$

and so on. Putting these into (3) and arranging terms in ascending order produces (2).

If we approximate f near (a, b) by cutting off the Taylor series (2) after the first-order terms, we have

$$f(x, y) \approx f(a, b) + f_x(a, b)(x - a) + f_y(a, b)(y - b). \tag{5}$$

That step is called **linearization** because the result is linear in x and y. Geometrically, (5) amounts to approximating the f surface, plotted above the x, y plane, by its **tangent plane** at (a, b) just as $f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots \approx f(a) + f'(a)(x-a)$ amounts to approximating the graph of f versus x by its **tangent line** at $x = a$ in the case of a function of a single variable.

Returning to (1), the idea is to expand P and Q about the singular point of interest, (x_s, y_s), and linearize. Using (5) to do that, we obtain

$$P(x,y) \approx P(x_s,y_s) + P_x(x_s,y_s)(x-x_s) + P_y(x_s,y_s)(y-y_s), \tag{6a}$$
$$Q(x,y) \approx Q(x_s,y_s) + Q_x(x_s,y_s)(x-x_s) + Q_y(x_s,y_s)(y-y_s). \tag{6b}$$

But $P(x_s, y_s)$ and $Q(x_s, y_s)$ are zero because (x_s, y_s) is a singular point, so we have the approximate (linearized) equations

$$\boxed{\begin{aligned} x' &= P_x(x_s,y_s)(x-x_s) + P_y(x_s,y_s)(y-y_s), \\ y' &= Q_x(x_s,y_s)(x-x_s) + Q_y(x_s,y_s)(y-y_s). \end{aligned}} \tag{7}$$

Finally, it is convenient, though not essential, to move the origin to S by letting $X \equiv x - x_s$ and $Y \equiv y - y_s$, and calling $P_x(x_s, y_s) = a$, $Q_x(x_s, y_s) = c$, $P_y(x_s, y_s) = b$, and $Q_y(x_s, y_s) = d$, for brevity, in which case (7) becomes

$$X' = aX + bY, \tag{8a}$$
$$Y' = cX + dY, \tag{8b}$$

which system is studied in Section 7.3. There, we classify the singularity at $X = Y = 0$ as a center, focus, node, or saddle, depending upon the roots of the characteristic equation

$$\lambda^2 - (a+d)\lambda + (ad - bc) = 0, \tag{9}$$

namely,

$$\lambda = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4bc}}{2}. \tag{10}$$

Knowing the numerical values of a, b, c, d, for the singular point in question, we can find the λ roots and determine whether the singular point is a center, focus, node, or saddle.

Understand that we are trying to ascertain the local behavior of the flow corresponding to the original nonlinear system (1) near the singular point S by studying the simpler linearized version of (1) at S, namely, (8). That program begs this question: Is the nature of the singularity of the nonlinear system (1), at S, truly captured by its linearized version (8)? To answer that question it is helpful to present the singularity classification, developed in Section 7.3, in a graphical form – as we have in Fig. 1. It is more convenient to deal with the two quantities $\mathrm{p} \equiv \mathrm{a} + \mathrm{d}$ and

$\mathbf{q \equiv ad - bc}$ (which are the axes in Fig. 1) than with the four quantities a, b, c, d since $a + d$ and $ad - bc$ determine the roots of (10) and hence the singularity type. In terms of p and q, (10) simplifies to $\lambda = (p \pm \sqrt{p^2 - 4q})/2$.

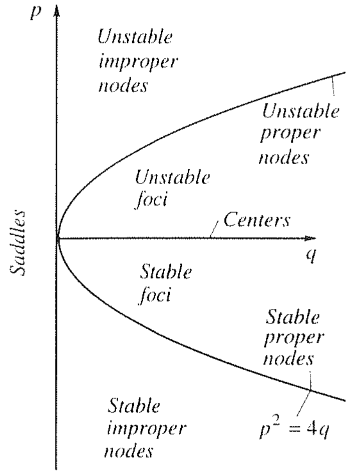


Figure 1. Singularity diagram; $p = a + d$, $q = ad - bc$.

In the figure there are five regions separated by the p axis, the parabola $p^2 = 4q$, and the positive q axis. [Of these boundaries, the p axis can be discounted since the case $q = 0$ is ruled out of consideration in Section 7.3 because then there is a line of singular points through the origin rather than the origin being an isolated singular point. That is, our (p, q) point will not fall on the boundary between saddles and improper nodes – namely, the p axis.]

The **Hartman–Grobman theorem** tells us that if a, b, c, d are such that the point (p, q) is *within* one of those regions, then the singularity types of the nonlinear system and its linearized version are identical. For instance, if the linearized system has a saddle, then so does the nonlinear system. Essentially, we can think of retaining the higher-order terms (which we drop in linearizing the nonlinear differential equations) as equivalent to infinitesimally perturbing the values of the coefficients a, b, c, d in the linearized equations and hence the values of p and q. If the point (p, q) is within one of the five regions, then the perturbed point will be within the same region, so the singularity type will be the same for the nonlinear system as for the linearized one.

However, we can imagine that for a borderline case, where (p, q) is on the parabola $p^2 = 4q$ or on the positive q axis, such a perturbation can push the point into one of the neighboring regions, thus changing the type. In fact, that is the way it turns out. For instance, if (p, q) is on the positive q axis (as occurs in the example to follow), then the nonlinear system could have an unstable focus or a center or a stable focus.

EXAMPLE 1. *Oscillator with Cubic Damping.* The equation $x'' + \epsilon x'^3 + x = 0$ models a harmonic oscillator with cubic damping – that is, with a damping term proportional to the velocity cubed. The equivalent system

$$\begin{aligned} x' &= y, \\ y' &= -x - \epsilon y^3 \end{aligned} \tag{11}$$

has one singular point, a center at $x = y = 0$. The linearized version is

$$\begin{aligned} X' &= Y = 0X + 1Y, \\ Y' &= -X = -1X + 0Y \end{aligned} \tag{12}$$

so $a = d = 0$, $b = 1$, and $c = -1$ hence $p = 0$ and $q = 1$. Thus, $(p, q) = (0, 1)$ so (from Fig. 1) the linearized system (12) has a center (no surprise, since the solutions of the linearized equation $x'' + x = 0$ are simple harmonic motions). However, it turns out (Exercise 1) that the nonlinear system (11) has a stable focus. ▮

To summarize, in general the linearized system faithfully captures the singularity type of the original nonlinear system. For the borderline cases, where (p, q)

is on the $p^2 = 4q$ parabola or on the positive q axis, however, we have these possibilities:

LINEARIZED		NONLINEAR
stable proper node	$\Longleftrightarrow$	stable focus, or stable proper node, or stable improper node
center	$\Longleftrightarrow$	unstable focus, or center, or stable focus
unstable proper node	$\Longleftrightarrow$	unstable improper node, or unstable proper node, or unstable focus

7.4.2. Applications. Consider some physical applications.

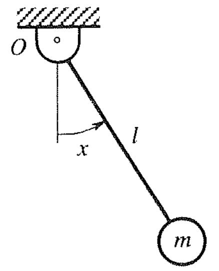


Figure 2. Pendulum.

EXAMPLE 2. *Pendulum.* Recall from an introductory physics course that for a rigid body undergoing pure rotation about a pivot axis the inertia about the pivot O times the angular acceleration is equal to the applied torque. For the pendulum shown in Fig. 2, the inertia about O is ml^2, the angle from the vertical is $x(t)$, the angular acceleration is $x''(t)$, and the downward gravitational force mg gives a torque of $-mgl \sin x$. If the air resistance is proportional to the velocity lx', say clx', then it gives an additional torque $-cl^2x'$, so the equation of motion is $ml^2x'' = -mgl \sin x - cl^2x'$ or

$$x'' + rx' + \frac{g}{l} \sin x = 0, \tag{13}$$

where $r \equiv c/m$. The rx' term is a damping term. For definiteness, let $g/l = 1$ and consider the undamped case, where $r = 0$.

For small motions we can approximate $\sin x$ by the first term of its Taylor series, $\sin x \approx x$, so that we have the simple harmonic oscillator equation $x'' + x = 0$ or

$$x' = y \tag{14a}$$

$$y' = -x; \tag{14b}$$

(14) has a center at $x = y = 0$, and its by now familiar phase portrait is shown in Fig. 3.

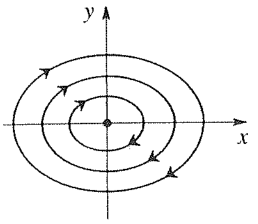


Figure 3. Phase portrait for the linearized system $x'' + x = 0$.

To study larger motions, suppose that we approximate $\sin x$ by the first *two* terms of its Taylor series instead: $\sin x \approx x - x^3/6$. Then we have the nonlinear, but still approximate, equation of motion

$$x'' + x - \frac{1}{6}x^3 = 0. \tag{15}$$

The latter is of the same form as the equation governing the rectilinear motion of a mass restrained by a "soft spring," which is studied in Section 7.2. The system

$$x' = y, \tag{16a}$$

$$y' = -x + \frac{1}{6}x^3 \tag{16b}$$

has a center at $(0,0)$ and saddles at $(\pm\sqrt{6}, 0)$ in the (x,y) phase plane, as discussed in Section 7.2, and its phase portrait is shown here in Fig. 4.

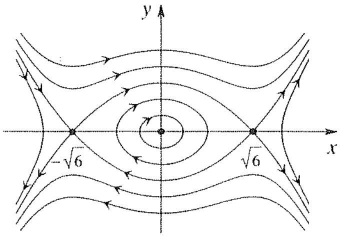


Figure 4. Phase portrait for the improved model (15).

Finally, if we retain the entire Taylor series of $\sin x$ (i.e., if we keep $\sin x$ intact), then we have the full nonlinear system (13), with $r = 0$, or

$$x' = y, \tag{17a}$$
$$y' = -\sin x, \tag{17b}$$

with singular points at $(x,y) = (n\pi, 0)$ for $n = 0, \pm 1, \pm 2, \ldots$. To classify these singularities, let us linearize equations (17) about the singular point $(n\pi, 0)$ using (7). Doing so, knowing that $\sin n\pi = 0$ and $\cos n\pi = (-1)^n$, and setting $X = x - n\pi$ and $Y = y - 0 = y$, the linearized version of (17) is

$$X' = Y = 0X + 1Y \tag{18a}$$
$$Y' = (-1)^{n+1}X = (-1)^{n+1}X + 0Y. \tag{18b}$$

In the notation of equation (8), $a = d = 0$, $b = 1$, and $c = (-1)^{n+1}$ so $p = a + d = 0$ and $q = ad - bc = (-1)^n$. Thus, these singular points are on the q axis in the p, q plane. For even integers n they are on the positive q axis and correspond to centers; for odd integers n they are on the negative q axis and correspond to saddles. In turn, the latter correspond to saddles of the nonlinear system (17), but the former could be centers or foci of (17), as is discussed in Section 7.4.1. The computer-generated phase portrait in Fig. 5 reveals that they are centers; we have centers at $x = 0, \pm 2\pi, \pm 4\pi, \ldots$, and saddles at $x = \pm\pi, \pm 3\pi, \ldots$, on the x axis.

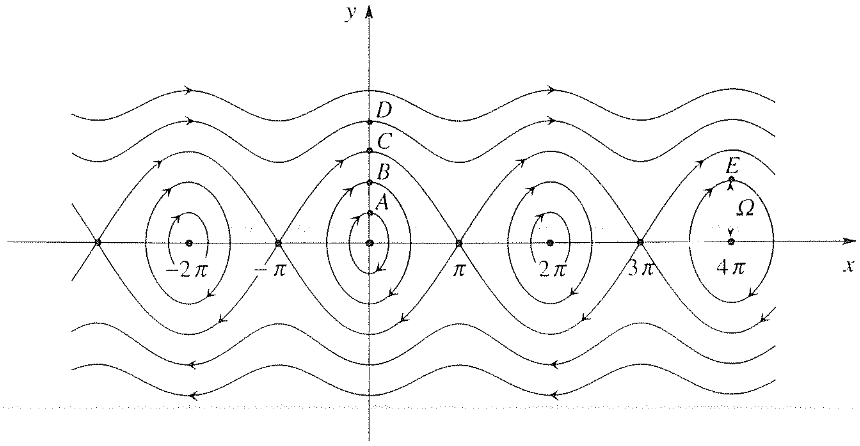


Figure 5. Phase portrait of the full nonlinear system $x'' + \sin x = 0$.

To understand the phase portrait, suppose (for definiteness) that the pendulum is hanging straight down ($x = 0$) initially, and that we impart an initial angular velocity $y(0)$, so that the initial point is A, B, C, or D in Fig. 5. If we start at A, then we follow a closed orbit that is very close to elliptical, and the motion is very close to simple harmonic motion at frequency $\omega = 1$. If we start at B, the orbit is not so elliptical, there is an increase in the period, and the motion deviates somewhat from simple harmonic. If we start at C, then we approach the saddle at $x = \pi$ as $t \to \infty$; that is, the pendulum approaches the inverted position as $t \to \infty$. If we impart more energy by starting at D, then – even though it slows down as it approaches the inverted position – it has enough energy to pass through that position and to keep going round and round indefinitely. Though the trajectory in the phase plane is not closed, the motion is nonetheless physically periodic since the positions x and $x + 2n\pi$ (for any integer n) are physically identical.

How can we gain access to one of the other closed orbits such as E? That's easy: we "crank" the pendulum through two rotations by hand so that while hanging straight down it is now at $x = 4\pi$. Then we impart an initial angular velocity $y(0) = \Omega$.

What is the equation of the trajectories? Dividing (17b) by (17a) and integrating, gives

$$\frac{1}{2}y^2 - \cos x = \text{constant} \equiv C. \tag{19}$$

Do we really need (19)? After all, we turned the phase portrait generation over to the computer. Yes, to help us choose initial conditions that will enable us to generate the separatrix (the trajectories through the saddles) on the computer. With $x = \pi$ and $y = 0$, we find that $C = 1$, so the equation of the separatrix is $y^2 = 2(1 + \cos x)$. Be careful, because the initial point $x = 0$ and $y = 2$ will not generate the entire separatrix, but only the segment through that point, from $x = -\pi$ to $x = \pi$. To generate the next segment we could use an initial point $x = 2\pi$ and $y = 2$, and so on.

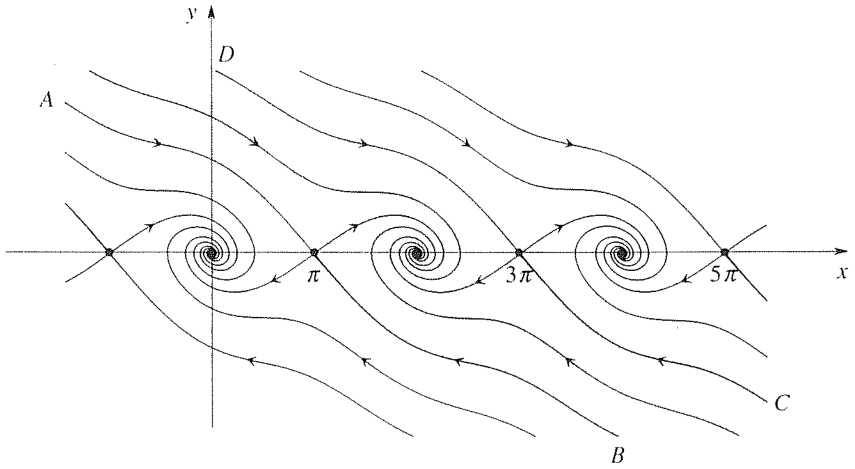


Figure 6. Phase portrait for the subcritically damped pendulum.

COMMENT 1. Recall that Fig. 3–5 correspond to taking $\sin x \approx x$, $\sin x \approx x - x^3/6$ and retaining $\sin x$ without approximation, respectively, in (13). Thus, and not surprisingly, as we retain more terms in the Taylor series approximation of $\sin x$ about $x = 0$ we capture the flow more accurately and completely.

COMMENT 2. What happens if we include some damping? It turns out that the singularities are at $(n\pi, 0)$, as before. If $r < r_{cr}$, where $r_{cr} = 2$, then the singularities are still saddles if n is odd, but if n is even we now have stable foci rather than centers as seen in Fig. 6 (for $r = 0.5$). One calls the lightly shaded region (not including the boundaries AB and CD) the **basin of attraction** for the stable focus at $(2\pi, 0)$, the basin of attraction of an attracting singular point S being the set of all initial points P_0 such that the representative point $P(t)$ tends to S as $t \to \infty$ if $P(0) = P_0$. Similarly, each of the other stable foci has its own basin of attraction.

COMMENT 3. We have spoken, in this section, of a nonlinear system having the same type of singularity (or not), at a particular singular point S, as the system linearized about S. Let us use the present example to clarify that idea. By their singularities being of the same type, we mean that their phase portraits are *topologically equivalent* in the neighborhood of S. Intuitively, that means that one can be obtained from the other by a continuous deformation, with the direction of the arrows preserved. The situation is illustrated in Fig. 7, where we show both the nonlinear (solid) and linearized (dashed) portraits in the neighborhood of the saddle at $(\pi, 0)$. [In more mathematical terms, suppose that our system $x' = P(x, y)$ and $y' = Q(x, y)$ has a singular point at the origin and that

$$P(x, y) = ax + by + \text{higher-order terms} \equiv aX + bY, \tag{20a}$$
$$Q(x, y) = cx + dy + \text{higher-order terms} \equiv cX + dY \tag{20b}$$

define X and Y as continuous functions of x and y, and vice versa. Such a relationship between x, y and X, Y is called a **homeomorphism** and is what we mean by a continuous deformation.]

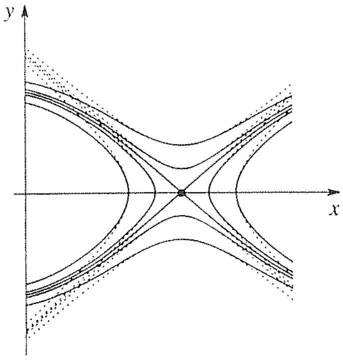


Figure 7. Continuous deformation of the portrait near $(\pi, 0)$. Solid lines correspond to the nonlinear system, dashed to the linearized version.

COMMENT 4. It turns out that the nonlinear pendulum equation is also prominent in connection with a superconducting device known as a *Josephson junction*. For discussion of the Josephson junction within the context of nonlinear dynamics, we recommend the book *Nonlinear Dynamics and Chaos* (Reading, MA: Addison–Wesley, 1994) by Steven H. Strogatz. ▮

In the preceding example we were unable to classify the singularities at $(n\pi, 0)$ with certainty, for n even and $r = 0$, as is discussed in the paragraph below (18). We relied on the computer-generated phase portrait, which show them to be centers, not foci. More satisfactorily, we could have used the fact that $x'' + \sin x = 0$ is a "conservative system," which idea we now explain.

In general, suppose that the application of Newton's second law gives

$$mx'' = F(x); \tag{21}$$

that is, where the force F happens to be an explicit function only of x, not of x' or t. Defining $V(x)$ by $F(x) = -V'(x)$, (21) becomes $mx'' + V'(x) = 0$. Let us multiply the latter by dx and integrate on x. Since, from the calculus,

$$x''dx = \frac{dx'}{dt}dx = \frac{dx'}{dt}\frac{dx}{dt}dt = x'\frac{dx'}{dt}dt = x'dx', \tag{22}$$

we obtain $mx'dx' + V'(x)dx = 0$, integration of which gives

$$\frac{1}{2}mx'^2 + V(x) = \text{constant}. \tag{23}$$

$V(x)$ is called the *potential energy* associated with the force $F(x)$. For a linear spring, for instance, the force is $F(x) = -kx$, and its potential is $V(x) = kx^2/2$. The upshot is that (23) tells us that the total energy (kinetic plus potential) is conserved, it remains constant over time, so we say that any system of the form (21) is **conservative**. The pendulum equation

$$ml^2x'' = -mgl\sin x \tag{24}$$

is of that form, and multiplying by dx and integrating on x gives

$$\frac{1}{2}m(lx')^2 - mgl\cos x = \text{constant}, \tag{25}$$

where $m(lx')^2/2$ is the kinetic energy and $-mgl\cos x$ is the potential energy associated with the gravitational force (with the pivot point chosen as the reference level).

For any conservative system (21), the total energy is $E(x,x') = mx'^2/2 + V(x)$. If we plot $E(x,x')$ above the x,x' phase plane, then the x,x' locations of maxima and minima of E are found from

$$\frac{\partial E}{\partial x} = V'(x) = 0 \qquad \text{and} \qquad \frac{\partial E}{\partial x'} = mx' = 0, \tag{26}$$

and these are precisely the singular points of the system

$$\begin{aligned} x' &= y, \\ y' &= F(x)/m = -V'(x)/m \end{aligned}$$

corresponding to (21). To illustrate, the point S beneath the minimum of E (Fig. 8) is a singular point. Furthermore, since E is constant on each trajectory, the phase trajectories are the projections of the closed curves of intersection of the E surface and the various horizontal planes, as sketched. Evidently (and we do not claim that this intuitive discussion constitutes a rigorous proof), a trajectory Γ very close to S must be a closed orbit. The only way Γ could fail to correspond to a periodic motion is if there is a singular point on Γ, for the flow would stop there. But if S is an isolated singular point, and Γ is small enough, then there can be no singular points on Γ, and we can conclude that S must be a center. By that reasoning, we could have known that the singularities at $(n\pi, 0)$ (for n even) must be centers, not foci. More generally, we state that *conservative systems do not have foci among their singularities.*

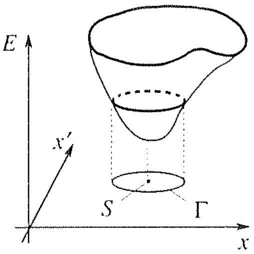


Figure 8. Occurrence of a center for a conservative system.

EXAMPLE 3. *Volterra's Predator-Prey Model.* The **Volterra model** (also known as the **Lotka–Volterra model**) of the ecological problem of two coexisting species, one the predator and the other its prey, is introduced in Section 7.2. Recall that if $x(t)$, $y(t)$ are the populations of prey and predator, respectively, then the governing equations are of the form

$$x' = \mu(1-y)x, \tag{27a}$$
$$y' = -\nu(1-x)y, \tag{27b}$$

where μ, ν are positive empirical constants. Setting the right-hand sides of (27) equal to zero reveals that there are two singular points: $(0,0)$ and $(1,1)$.

Linearizing (27) about $(0,0)$ gives

$$x' = \mu x, \qquad y' = -\nu y \tag{28}$$

so $a = \mu$, $b = c = 0$, $d = -\nu$. Hence, $\lambda = \mu$ and $-\nu$, which are of opposite sign, so the singularity at the origin is a saddle. Clearly, the straight-line trajectories through $(0,0)$ are simply the x and y axes since $x = 0$, $y = Ae^{-\nu t}$, and $x = Be^{\mu t}$, $y = 0$ satisfy (28) and give trajectories that pass through the origin.

To linearize about $(1,1)$ we use (7) and obtain the approximations

$$x' = -\mu(y-1), \qquad y' = \nu(x-1) \tag{29}$$

or, with $X \equiv x - 1$ and $Y \equiv y - 1$,

$$X' = -\mu Y = 0X - \mu Y, \tag{30a}$$

$$Y' = \nu X = \nu X + 0Y. \tag{30b}$$

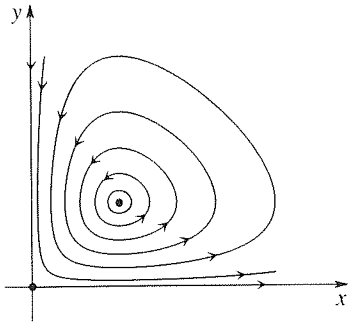


Figure 9. Phase portrait for Volterra problem (27).

Thus, $a = d = 0$, $b = -\mu$, $c = \nu$ so $\lambda = \pm 2i\sqrt{\mu\nu}$. Hence, the linearized version (30) has a center, and (27) has either a center or a focus.

The phase portrait in Fig. 9 shows the singularity at $(1,1)$ to be a center, with every trajectory being a periodic orbit, except for the two coordinate axes. (Of course, we show only the first quadrant because x and y are populations and hence nonnegative.) The direction of the arrows follows from (27), which reveals that $x' > 0$ for $y < 1$, and $x' < 0$ for $y > 1$ (or, $y' < 0$ for $x < 1$ and $y' > 0$ for $x > 1$).

COMMENT. Although the Volterra model is a useful starting point in the modeling process and is useful pedagogically, it is not regarded as sufficiently realistic for practical ecological applications. ▮

7.4.3. Bifurcations. As we have stressed, our approach in this chapter is largely qualitative. Of special importance, then, is the concept of bifurcations. That is, systems generally include one or more physical parameters (such as μ and ν in Example 3). As those parameters are varied continuously, one expects the system behavior to change continuously as well. For instance, if we vary μ in Example 3, then the eccentricity of the orbits close to the center at $(1,1)$ changes, and the overall flow field deforms, but – qualitatively – nothing dramatic happens. In other cases, there may exist certain critical values of one or more of the parameters such that the overall system behavior changes abruptly and dramatically as a parameter passes through such a critical value. We speak of such a result as a **bifurcation**. Let us illustrate the idea with an example.

EXAMPLE 4. *Saddle-Node Bifurcation.* The nonlinear system

$$x' = -rx + y, \tag{31a}$$

$$y' = \frac{x^2}{1+x^2} - y \tag{31b}$$

arises in molecular biology, where $x(t)$ and $y(t)$ are proportional to protein and messenger RNA concentrations, and r is a positive empirical constant, or parameter, associated with the "death rate" of protein in the absence of the messenger RNA [for if $y = 0$, then (31a) gives exponential decay of x, with rate constant r].

The singular points of (31) correspond to intersection points of $y = rx$ and $y = x^2/(1+x^2)$, as shown (solid curves) in Fig. 10. Equating these gives $x = y = 0$ and also the two distinct roots

$$x_{\pm} = \frac{1 \pm \sqrt{1-4r^2}}{2r}, \qquad y_{\pm} = rx_{\pm} = \frac{1 \pm \sqrt{1-4r^2}}{2}, \tag{32}$$

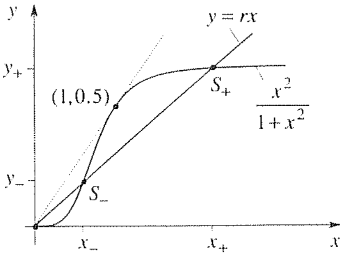


Figure 10. Determining the singular points of (31).

provided that $r < 1/2$. Thus, the critical slope of $y = rx$ is $r = 1/2$. If $r < 1/2$ we obtain the two intersections $S_+ \equiv (x_+, y_+)$ and $S_- \equiv (x_-, y_-)$ if $r = 1/2$ (dashed line in Fig. 10) these coalesce at $(1, 0.5)$, and if $r > 1/2$ they disappear and we have only the singular point at the origin.

Let us study the three singular points, for $r < 1/2$. First $(0,0)$: we can see from (31) by inspection or Taylor series expansion, that the linearized equations are

$$x' = -rx + y, \qquad y' = -y \tag{33}$$

so $a = -r$, $b = 1$, $c = 0$, and $d = -1$. Thus, (10) gives $\lambda = -r$ and -1. Since both are negative, the singular point $(0,0)$ is a stable node.

In similar fashion (which calculations we leave to Exercise 6), we find that the singularity at S_- is a saddle, and that the singularity at S_+ is an unstable improper node. As r is increased, S_- and at S_+ approach each other along the curve $y = x^2/(1+x^2)$. When $r = 1/2$ they merge and form a singularity of some other type, and when r is increased beyond $1/2$ the singularity disappears altogether, leaving only the node at the origin. The bifurcation that occurs at $r = 1/2$ is an example of a "saddle-node bifurcation." From the way the singular points S_+ and S_- approach each other along the unstable manifold of the saddle, like "beads on a string" as Strogatz puts it, we see that the bifurcation process is essentially a one-dimensional event embedded within a higher-dimensional space (two-dimensional in this case). ▮

The saddle-node bifurcation illustrated above is but one type of bifurcation. A few others are discussed in the exercises and in the next section. For a more complete discussion of bifurcation theory, we recommend the book by Strogatz, referenced in Example 2.

Closure. In this section we got into the details of the phase plane analysis of autonomous nonlinear systems. Whether or not we generate the phase portrait by computer, it is essential to begin an analysis by finding any singular points and, by linearization, to determine the key features of the local flow near each singular point. That information is needed even if we turn to the computer to generate the phase portrait, as we discuss below, under "Computer software."

We also explored the correspondence between the type of a singularity of the nonlinear system and that of the linearized system and found that the type remains the same, except for the borderline cases corresponding to p, q points on the positive q axis or on the parabola $p^2 = 4q$ in Fig. 1. Those cases could "go either way." That is, the higher-order terms in the nonlinear system can be thought of as perturbing the a, b, c, d coefficients in the linearized equations and hence p, q as well. The perturbed p, q point could remain on the boundary curve (positive q axis or the parabola $p^2 = 4q$), or it could be pushed slightly into either of the adjoining regions, thus changing the singularity type.

The other item of special importance is the notion of bifurcations, whereby dramatic and qualitative changes in the system behavior can result from a given parameter crossing a critical value known as a bifurcation point.

Computer software. Consider, for instance, the computer generation of the trajectories through the singular point $(\pi, 0)$ in Fig. 6. The idea is to express the governing equation $x'' + 0.5x' + \sin x = 0$ (we took $r = 0.5$ in Fig. 6) as the system

$$\begin{aligned} x' &= y, \\ y' &= -\sin x - 0.5y \end{aligned} \tag{34}$$

and to linearize (34) about $(\pi, 0)$ as

$$\begin{aligned} x' &= y, \\ y' &= -(\cos\pi)(x-\pi) - 0.5y \end{aligned} \tag{35}$$

or with $x - \pi \equiv X$ and $y - 0 \equiv Y$,

$$\begin{aligned} X' &= 0X + 1Y, \\ Y' &= 1X - 0.5Y. \end{aligned} \tag{36}$$

Next, seek straight-line solutions as $Y = \kappa X$, so (36) becomes

$$\begin{aligned} X' &= 0X + \kappa X, \\ \kappa X' &= 1X - 0.5\kappa X, \end{aligned} \tag{37}$$

comparison of which gives $0 + \kappa = (1 - 0.5\kappa)/\kappa$ and hence $\kappa = -1.2807764$ and 0.7807764. Thus, if we make very small steps along the line $Y = -1.2807764X$, away from $(\pi, 0)$, we can obtain initial points that are very close to being on the trajectories from A to $(\pi, 0)$ and from B to $(\pi, 0)$. For example, with $X = -0.02$ we obtain $Y = 0.02561553$. Similarly, $X = +0.02$ gives $x = 3.1615927$ and $y = -0.02561553$.

With these two initial points we can use the *Maple* phaseportrait command to obtain the desired trajectories from A to $(\pi, 0)$ and from B to $(\pi, 0)$, as follows. Remembering to first enter with(DEtools):, the phaseportrait command would be as follows:

```
phaseportrait([y, -.5 * y - sin (x)], [t, x, y], t = -10..10,
{[0, 3.1215927, .02561553], [0, 3.1615927, -.02561553]}, stepsize = .05,
x = -4.5..17, y = -6..6, scene = [x, y]);
```

The t range, from -10 to $+10$, is found to be sufficient to give the desired trajectories all the way to the borders $y = -6$ and $y = +6$, and the x and y ranges are simply chosen as the same ones used in Fig. 6.

EXERCISES 7.4

1. (a) In Example 1 we stated that the equation $x'' + \epsilon x'^3 + x = 0$ $(\epsilon > 0)$ has a stable focus at the origin of the phase plane. Verify that claim by generating (with computer software) its phase portrait.
(b) Should the cubic damping result in the oscillation dying out more or less rapidly than for the case of linear damping, $x'' + \epsilon x' + x = 0$, for the same values of ϵ? Explain.
(c) Classify the singularity at $(0, 0)$ for the case where $\epsilon < 0$, and support your claim.

2. Determine all singular points, if any, and classify each insofar as possible.

(a) $x' = y, \quad y' = 1 - x^4$
(b) $x' = 1 - y^2, \quad y' = 1 - x$
(c) $x' = y, \quad y' = (1 - x^2)/(1 + x^2)$
(d) $x' = x - y, \quad y' = \sin(x + y)$
(e) $x' = (1 - x^2)y, \quad y' = -x - 2y$
(f) $x' = (1 - x^2)y, \quad y' = -x + 2y$
(g) $x' = -2x - y, \quad y' = x + x^3$
(h) $x' = -2x - y, \quad y' = \sin x$
(i) $x' = ye^x - 1, \quad y' = y - x - 1$
(j) $x' = x^2 - 2y, \quad y' = 2x - y$
(k) $x' = x^2 - y^2, \quad y' = x^2 + y - 2$
(l) $x' = y, \quad y' = -3\sin x$
(m) $x' = x + 2y, \quad y' = -x - \sin y$
(n) $x' = (x^6 + 1)y, \quad y' = x^2 - 4$

3. (a)–(n) Use computer software such as the *Maple* phaseportrait command, to generate the phase portrait of the corresponding system in Exercise 2.

4. Is the given system conservative? Explain.

(a) $x'' - 2x' + \sin x = 0$ (b) $x'' + x'^2 + x^2 = 0$
(c) $x'' + x^2 = 0$ (d) $x'' + x' + x = 0$

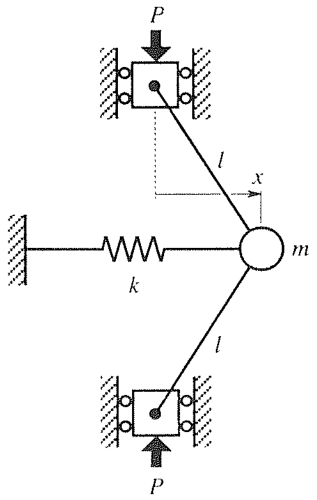
5. Use computer software such as the *Maple* phaseportrait command, to obtain the phase portrait of equation (13), over $-2 < x < 14$ and $-3 < y < 3$, showing enough trajectories to clearly portray the flow. Take $g/l = 1$, and

(a) $r = 0$.
(b) r to be any positive value that you wish but small enough for the damping to be subcritical. For instance, the singularity at $(0, 0)$ should be a stable focus.
(c) r to be any value that you wish but large enough for the damping to be supercritical. For instance, the singularity at $(0, 0)$ should be a stable node.

6. (*Example 4 continuation.*) In parts (a) and (b) below, let $r = 0.3$ in (31).

(a) Show that S_- is a saddle, and find the equations of the two straight-line trajectories through it. Show that S_+ is a stable improper node, and find two straight-line trajectories through it. Find two straight-line trajectories through the stable node at $(0, 0)$. Use these results to sketch the phase portrait of (31).
(b) For the critical case, $r = 1/2$, show that the singularity at $S_\pm$ is nonelementary.
(c) For the representative supercritical case $r = 1$, identify and classify any singularities, and use computer software to generate the phaseportrait of (31) in the rectangle $0 \le x \le 1.5$, $0 \le y \le 1.5$.

7. (*Dynamic formulation of a buckling problem*) Consider the buckling of the mechanical system shown in the figure, and consisting of two massless rigid rods of length l pinned to a mass m and a lateral spring of stiffness k. That is, when the spring is neither stretched nor compressed $x = 0$ and the rods



are aligned vertically. As we increase the downward load P nothing happens until we reach a critical value P_{cr}, at which value x increases (to one side or the other, we can't predict which) and the system collapses.

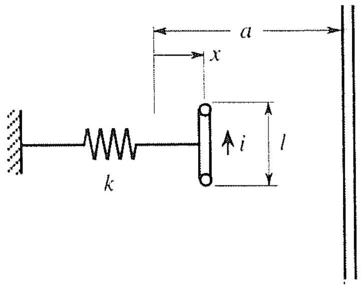
(a) Application of Newton's second law of motion gives

$$mx'' - \frac{2Px}{l}\left[1-\left(\frac{x}{l}\right)^2\right]^{-1/2} + kx = 0 \qquad (7.1)$$

as governing the displacement $x(t)$. With $x' = y$, show that the singularity at the origin in the x, y phase plane changes its type as P is sufficiently increased. Discuss that change of type, show how it corresponds to the onset of buckling, and use it to show that the critical buckling load is $P_{cr} = kl/2$.
(b) Explain what the results of part (a) have to do with bifurcation theory.
(c) Use Newton's law to derive (7.1).

8. (*Motion of current-carrying wire*) A mutual force of attraction is exerted between parallel current-carrying wires. The infinite wire shown in the figure has current I, and the wire of length l and mass m (with leads that are perpendicular to the paper) has current i in the same direction as I. According to the Biot-Savart law, the mutual force of attraction is $2Iil/(\text{separation}) = 2Iil/(a-x)$, where $x = 0$ is the position at which the spring force is zero, so the equation of motion of the restrained wire is

$$mx'' + k\left(x - \frac{r}{a-x}\right) = 0, \quad \text{where} \quad r = \frac{2Iil}{k}. \qquad (8.1)$$



Thinking of m, k, a, and l as fixed, and the currents I and i as variable, let us study the behavior of the system in terms of the parameter r. For definiteness, let $m = k = a = 1$.

(a) With $x' = y$, identify any singularities in the x, y phase plane and their types, and show that they depend upon whether r is less than, equal to, or greater than $1/4$. Suppose that $r < 1/4$. Find the equation of the phase trajectories and of the separatrix. Do a labeled sketch of the phase portrait.
(b) Let $r = 0.1$, say, and obtain a computer plot of the phase portrait.
(c) Next, consider the transitional case, where $r = 1/4$. Show that that case corresponds to the merging of the two singularities, and the forming of a single singularity of higher order (i.e., a nonelementary singularity). Do a labeled sketch of the phase portrait for that case.
(d) Let $r = 1/4$, and obtain a computer plot of the phase portrait.
(e) Next, consider the case where $r > 1/4$, and sketch the phase portrait.
(f) Let $r = 0.5$, say, and obtain a computer plot of the phase portrait.
(g) Discuss this problem from the point of view of bifurcations, insofar as the parameter r is concerned.

7.5 Limit Cycles, van der Pol Equation, and the Nerve Impulse

7.5.1. Limit cycles and the van der Pol equation. The **van der Pol equation**,

$$\boxed{x'' - \epsilon(1-x^2)x' + x = 0,} \qquad (\epsilon > 0) \qquad (1)$$

was studied by *Balth van der Pol* (1889–1959), first in connection with current oscillations in a certain vacuum tube circuit and then in connection with the modeling of the beating of the heart.* Usually, in applications the parameter ϵ is positive.

To study (1) in the phase plane we first re-express it as the system

$$x' = y, \tag{2a}$$
$$y' = -x + \epsilon(1 - x^2)y, \tag{2b}$$

which has one singular point: $(0, 0)$. Linearizing (2) about $(0, 0)$ gives

$$x' = y, \tag{3a}$$
$$y' = -x + \epsilon y, \tag{3b}$$

which has an unstable focus if $\epsilon < 2$ and an unstable node if $\epsilon > 2$. That result is not surprising since (3) is equivalent to $x'' - \epsilon x' + x = 0$ [equation (1) with the nonlinear $\epsilon x^2 x'$ term dropped], and the latter corresponds to a damped harmonic oscillator with *negative* damping. Near the origin in the x, y phase plane the flow is accurately described by (3) and is shown in Fig. 1a. As the motion increases, the neglected nonlinear term $\epsilon x^2 x'$ ceases to be negligible, and we wonder how the trajectory shown in Fig. 1a continues to develop as t increases. Since the "damping coefficient" $c = -\epsilon(1-x^2)$, in (1), is negative throughout the vertical strip $|x| < 1$, we expect the spiral to continue to grow, with distortion as the $\epsilon x^2 x'$ term becomes more prominent. Eventually, the spiral will break out of the $|x| < 1$ strip (Fig. 1b). As the representative point $(x(t), y(t))$ spends more and more time outside that strip, where $c = -\epsilon(1 - x^2) > 0$, the effect of the positive damping in $|x| > 1$ increases, relative to the effect of the negative damping in $|x| < 1$, so it is natural to wonder if the trajectory might approach some limiting closed orbit as $t \to \infty$ over which the effects of the positive and negative damping are exactly in balance.

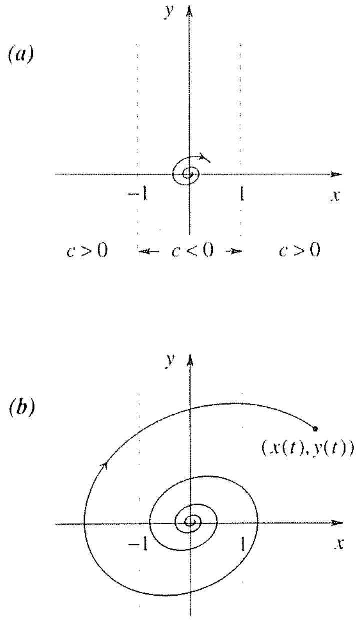


Figure 1. The unstable focus at$(0, 0)$.

We can use the following theorem, due to N. Levinson and O. K. Smith.

THEOREM 7.5.1 *Existence of Limit Cycle*
Let $f(x)$ be even $[f(-x) = f(x)]$ and continuous for all x. Let $g(x)$ be odd $[g(-x) = -g(x)]$ with $g(x) > 0$ for all $x > 0$, and $g'(x)$ be continuous for all x. With

$$\int_0^x f(\xi)\, d\xi \equiv F(x) \quad \text{and} \quad \int_0^x g(\xi)\, d\xi \equiv G(x), \tag{4}$$

suppose that (i) $G(x) \to \infty$ as $x \to \infty$ and (ii) there is an $x_0 > 0$ such that $F(x) < 0$ for $0 < x < x_0$, $F(x) > 0$ for $x > x_0$, and $F(x)$ is monotonically increasing for $x > x_0$ with $F(x) \to \infty$ as $x \to \infty$. Then the *generalized Liénard equation*

$$x'' + f(x)x' + g(x) = 0 \tag{5}$$

has a single periodic solution, the trajectory of which is a closed curve encircling the origin in the x, x' phase plane. All other trajectories (except the trajectory

[*] B. van der Pol, On "Relaxation Oscillations," *Philosophical Magazine*, Vol. 2, 1926, pp. 978–992, and B. van der Pol and J. van der Mark, The Heartbeat As a Relaxation Oscillation, and An Electrical Model of the Heart, *Philosophical Magazine*, Vol.6, 1928, pp. 763–775.

consisting of the single point at the origin) spiral toward the closed trajectory as $t \to \infty$.

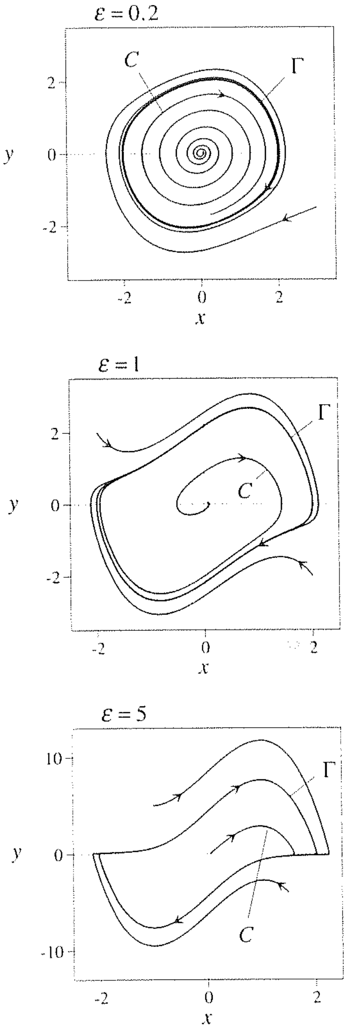


Figure 2. The van der Pol limit cycle, for $\epsilon = 0.2, 1,$ and 5.

Applying this theorem to the van der Pol equation (1), $f(x) = -\epsilon(1 - x^2)$ is an even function of x and $F(x) = -\epsilon(x - x^3/3)$, which is less than zero for $0 < x < \sqrt{3}$, greater than zero for $x > \sqrt{3}$, and which increases monotonely to infinity as $x \to \infty$. Further, $g(x) = x$ is odd, and positive for $x > 0$, $g'(x) = 1$, and $G(x) = x^2/2 \to \infty$ as $x \to \infty$. Since the conditions of the theorem are met (for all $\epsilon > 0$), we conclude from the theorem that the van der Pol equation does admit a closed trajectory, a periodic solution, for every positive value of ϵ.

Computer results (using the *Maple* phaseportrait command) bear out this claim. The phase portraits are shown in Fig. 2 for the representative cases $\epsilon = 0.2, 1,$ and 5, and $x(t)$ is plotted versus t in Fig. 3 for the trajectories labeled C. The closed trajectories labeled Γ, predicted by the theorem, are examples of **limit cycles** – namely, isolated closed orbits. By Γ being isolated we mean that neighboring trajectories through points arbitrarily close to Γ are not closed orbits. If we start on Γ we remain on Γ, but if we start on a neighboring trajectory, then we approach Γ as $t \to \infty$ (unless we start at the origin, which is an equilibrium point). Thus, we classify the van der Pol limit cycle as **stable** (or **attracting**). Clearly, that particular trajectory is of the greatest importance because every other trajectory (except the point trajectory $x = y = 0$) winds onto it as $t \to \infty$.

As one might suspect from Fig. 2 and as can be proved, the van der Pol limit cycle approaches a circle of radius 2 as $\epsilon \to 0$ through positive values. When ϵ becomes zero the singularity at the origin changes from a focus to a center, and while the circle of radius 2 persists as a trajectory it is joined by the whole family of circular orbits centered at the origin. If ϵ is diminished further and becomes negative, the origin becomes a stable focus and all closed orbits disappear and give way to inward-winding spirals. Thus, $\epsilon = 0$ is a bifurcation value of ϵ.

Observe the interesting extremes: as $\epsilon \to 0$, the steady-state oscillation (i.e., corresponding to the limit cycle) becomes a purely harmonic motion with amplitude 2. But as ϵ becomes large, the limit cycle distorts considerably and the steady-state oscillation $x(t)$ becomes "herky jerky." (In the exercises, we show that as $\epsilon \to \infty$ it even becomes discontinuous!) Such motions were dubbed as **relaxation oscillations** by van der Pol, and these are found all around us. Just a few, mentioned in the paper by van der Pol and van der Mark, are the singing of wires in a cross wind, the scratching noise of a knife on a plate, the squeaking of a door, the intermittent discharge of a capacitor through a neon tube, the periodic reoccurrence of epidemics and economic crises, the sleeping of flowers, menstruation, and the beating of the heart. Such oscillations are characterized by a slow buildup followed by a rapid discharge, then a slow buildup, and so on. Thus, there are two time scales present, a "slow time" associated with the slow buildup, and a "fast time" associated with the rapid discharge. In biological oscillators such as the heart, the period of oscillation provides a biological "clock."

Understand clearly that the *limit cycle phenomenon is possible only in nonlinear systems* for consider the case of small ϵ, say, where we have a limit cycle that is approximately a circle of radius 2. If the system were linear, then the existence of that orbit would imply that the entire family of concentric circles would necessarily be trajectories as well, but they are not.

Besides the van der Pol example, other examples of differential equations exhibiting limit cycles are given in the exercises. In other cases a limit cycle can be **unstable (repelling)** in that other trajectories wind away from it, or **semistable** in exceptional cases, in that other trajectories wind toward it from the interior and away from it on the exterior, or vice versa.

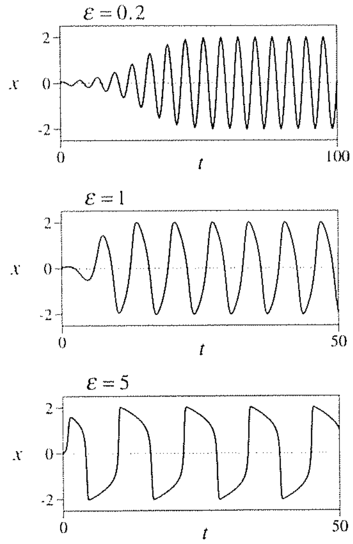


Figure 3. $x(t)$ versus t.

7.5.2. Application to the nerve impulse and visual perception. The brain contains about 10^{12} (a million million) neurons, with around 10^{14} to 10^{15} interconnections. Within this complex network, information is encoded and transmitted in the form of electrical impulses. The basic building block is the individual neuron, or nerve cell, and the functioning of a single neuron as an input/output device is of deep importance and interest. Our purposes in discussing the neuron here are in connection with relaxation oscillations, and especially with the key role of nonlinearity in the design and functioning of our central nervous system.

A typical neuron is comprised of a cell body that contains the nucleus and that emanates many dendrites and a single axon. The axon splits near its end into a number of teminals as shown schematically in Fig. 4. Dendrites are on the order of a millimeter long, and axons can be as short as that or as long as a meter. At the end of each terminal is a synapse, which is separated from a dendrite of an adjacent cell by a tiny synaptic gap. Electrical impulses, each of which is called an **action potential**, are generated near the cell body and travel down the axon. When an action potential arrives at a synapse, chemical signals in the form of neurotransmitter molecules are released into the synaptic gap and diffuse across that gap to a neighboring dendrite. These electrical signals to that neighboring neuron can be positive (excitatory) or negative (inhibitory). Each cell receives a great many such cues from other neurons. If the net excitation to a cell falls below some critical threshold value, then the cell will not fire – that is, it will not generate action potentials. If the net excitation is a bit above that threshold, then the cell will fire – not just once, but repeatedly and at a certain frequency.

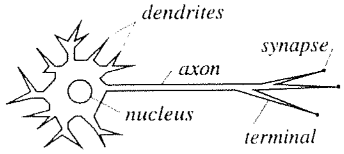


Figure 4. Typical neuron.

Let us consider briefly the generation of the nerve impulse. The nerve cell is surrounded by and also contains salt water. The salt molecules include sodium chloride (NaCl) and potassium chloride (KCl), and many of these molecules are ionized so that Na^+, K^+, and Cl^- ions are abundant both inside and outside the axon. Of these, Na^+ and K^+ are the key players insofar as the nerve impulse is concerned. Rather than being impermeable, the axon membrane has many tubular protein pores of two kinds: channels that can open or close and let either Na^+ or K^+ ions through in a passive manner, like valves, and pumps that (using energy from the metabolism of glucose and oxygen) actively eject Na^+ ions (i.e., from inside the axon to outside) and bring in K^+ ions. Through the action of these active and passive pores, and the physical mechanisms of diffusion and the repulsion/attraction of like/unlike charges, a differential in charge, and hence potential (voltage), is established across the axon membrane which, in the resting state, is 70 millivolts, positive outside.

If the net excitation arriving from other cells sufficiently reduces that voltage, at the cell body end of the axon, then a sequence of opening and closing of pores is established, which results in a flow of Na^+ and K^+ ions and hence a voltage "blip," the action potential, proceeding down the axon. That wave is not like the flow of electrons in a copper wire, but rather like a water wave that results not from horizontal motion of the water, but from a differential up and down motion of water particles. This complicated process was pieced together by Alan Hodgkin and Andrew Huxley, in 1952, and is clearly discussed in the book by David H. Hubel.* Various electrical circuit analogs have been proposed, to model the nerve impluse, by Hodgkin and Huxley and others. They are all somewhat empirical and of the "fill and flush" type – that is, where a charge builds up and is then discharged through an electrical circuit, and they are described in the little book by F. C. Hoppensteadt.†

Of interest to us here is that the firing is repetitive (on the order of 100 impulses per second), and consists of a relaxation oscillation governed by the van der Pol equation (as discussed in Hoppensteadt). Further, it is known that as the excitation voltage is increased above the threshold, the magnitude of the action potential remains unchanged, but the firing frequency increases. If we plot the output (action potential) amplitude versus the input (excitation voltage) amplitude, the graph is as shown in Fig. 5. Since the graph of output amplitude versus input amplitude is not a straight line through the origin, the process must be nonlinear, which fact is also known through the governing equation being a van der Pol equation (or other such equation, depending upon the model adopted); indeed, any process where the output amplitude is zero until a critical threshold is reached is necessarily nonlinear. Since the individual neuron is a nonlinear device, surely the same is true for the entire central nervous system, and the natural and important question that asserts itself is "Why?". What is the functional purpose of that nonlinearity?

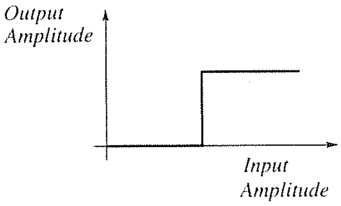


Figure 5. Input/output relation for a neuron.

Let us attempt an answer. We have seen that nonlinear systems are more complex than linear ones. Since our nervous system is responsible for carrying out complex tasks, it seems reasonable that the system chosen should be nonlinear. We can be more specific if we look at a single type of task, say visual perception, which is so complex that it occupies around a third of the million million neurons in the brain.

Perhaps the most striking revelation in studying visual perception is in discovering that one's visual perception is not a simple replica of the image that falls upon the retina but is an interpretation of that information, effected by visual processing that begins in the retina and continues up into the visual cortex of the brain. For instance, hold your two hands up, in front of your face, with one twice as far from your eyes as the other (about 8 and 16 inches). You should find that they look about the same size. Yet, if we replace your eyes with a camera, and take a picture, we see in the photo that one hand looks around twice as large as the other. Usually, we blame the camera for the "distortion," but the camera simply shows you the same

**Eye, Brain, and Vision* (New York: W. H. Freeman and Company, 1988).

†*An Introduction to the Mathematics of Neurons* (Cambridge: Cambridge University Press, 1986).

information that is picked up by your retinas, the distortion is introduced by the brain as it interprets and reconstructs the data before presenting it to you as visual consciousness.

The latter is but one example of a principle of visual perception known as **size constancy**. The idea is that whereas the actual size of a physical object is invariant, the size of its retinal image varies dramatically as it is moved nearer or further from us. Size constancy is the processing, between the retina and visual consciousness, that compensates for such variation so as to stabilize our visual world. Thus, for instance, our hands look about the same size even when the retinal image of one is twice as large as that of the other. The functional advantage of that stabilization is to relieve our conscious mind of having to figure everything out; our brain does much of the figuring out and presents us with its findings so that our conscious attention can be directed to more pressing and singular matters.

Surely, size constancy requires a nonlinear perceptual system for if we take the retinal image size as the input amplitude and the perceived size as the output amplitude, then a linear system would show us the two hands just as a camera does. (Remember that for a linear system if we double the input we double the corresponding output if we triple the input we triple the output, and so on.)

In visual perception there are other constancy mechanisms as well, such as **brightness constancy** and **hue constancy**. To illustrate brightness constancy, consider the following simple experiment reported by Hubel in his book, cited above. We know from experience that a newspaper appears pretty much the same, whether we look at it in sunlight or in a dimly lit room: black print on white paper. Taking a newspaper and a light meter, Hubel found that the white paper reflected 20 times as much light outdoors as indoors, yet it looked about the same outdoors and indoors. If the perceptual system were linear, the white paper should have *looked* 20 times as bright outdoors compared to indoors. Even more striking, he found that the black letters actually reflected twice as much light outdoors as the white paper did indoors yet, whether indoors or outdoors, the black letters always looked black and white paper always looked white.

The point, then, is that these constancy mechanisms stabilize our perceived world and relieve our conscious mind from having to deal with newspapers that look 20 times brighter outdoors than indoors, hands that "grow" and "shrink" as they are moved to and fro, and so on, so that our conscious attention can be reserved for more singular matters such as not getting hit by a bus. These mechanisms are possible only by virtue of the nonlinearity of the central nervous system, which can be traced, in turn, to the highly nonlinear behavior of the basic building block of that system, the individual neuron.

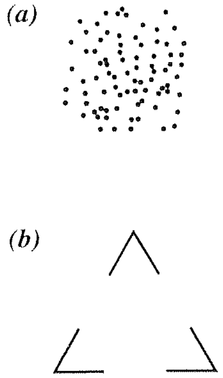


Figure 6. The whole is greater than the sum of its parts.

You have no doubt heard about "the whole being greater than the sum of its parts." That idea expresses the essence of the Gestalt school of psychology which, by around 1920, supplanted the previously dominant molecular school of psychology, which had held that the whole *is* equal to the sum of the parts. To illustrate the Gestalt view, notice that the black dots in Fig. 6a are seen as a *group* of dots, not as a number of individual dots, and that the arrangement in Fig. 6b is seen as a triangle with sections removed, rather than as three bent lines. In fact, Max Wertheimer's fundamental experiment, which launched the Gestalt concept in 1912, is as follows. If parallel lines of equal length are displayed on a screen successively, it is found that if the time interval and distance between them are sufficiently small, then they are perceived not as two separate lines but as a single line that *moves* laterally. (Today we recognize that idea as the basis of motion pictures!)

In mathematical terms, the molecular idea is reminiscent of the result for a linear differential equation $L[u] = f_1 + f_2 + \cdots + f_k$ that the response u to the combined input is simply the sum of the responses $u_1, u_2, \ldots, u_k$ to the individual inputs $f_1, f_2, \ldots, f_k$. We suggest here that, in effect, the contribution of the Gestaltists was to recognize the highly nonlinear nature of the perceptual system, even if they did not think of it or express it in those terms. We say more about the far-reaching effects of that nonlinearity upon human behavior in the next section.

Closure. The principal idea of this section is that of limit cycles, which occur only for nonlinear systems. The classic example of an equation with a limit cycle solution is the van der Pol equation, which we discuss, but that is by no means the only equation that exhibits a limit cycle. That limit cycle solution is said to be a **self-excited oscillation** because even the slightest disturbance from the equilibrium point at the origin results in an oscillation that grows and inevitably approaches the limit cycle as $t \to \infty$. The case of large ϵ is especially important in biological applications, and the corresponding limit cycle solution is a relaxation oscillation characterized by alternate t-intervals of slow and rapid change. Since the existence of a limit cycle is of great importance, there are numerous theorems available that help one to detect whether or not a limit cycle is present, but we include only the theorem of Levinson and Smith since it is helpful in our discussion of the van der Pol equation.

Finally, we discuss the action potential occuring during the firing of a neuron, as a biological illustration of a relaxation oscillation, and we use that example to point out the nonlinear nature of the neuron and central nervous system, and the profound implications of that nonlinearity.

EXERCISES 7.5

1. (a) Obtain computer results analogous to those presented in Fig. 2 and 3 for the case where $\epsilon = 0.1$. What value do you think the period approaches as $\epsilon \to 0$? Explain.
(b) Obtain computer results analogous to those presented in Fig. 2 and 3, for $\epsilon = 10$. NOTE: Be sure to make your t-integration step size small enough – namely, small compared to the time intervals of rapid change of $x(t)$.

2. Use Theorem 7.5.1 to show that the following equations admit limit cycles.

(a) $x'' - (1 - x^2)x' + x^3 = 0$
(b) $x'' + x^2(5x^2 - 3)x' + x = 0$

3. Identify and classify any singularities of the given equation in the x, y phase plane, where $x' = y$. Argue as convincingly as you can for or against the existence of any limit cycles, their shape, and their stability or instability. You should be able to tell a great deal from the equation itself, even in advance of any computer simulation.

(a) $x'' + (x^2 + x'^2 - 1)x' + x = 0$
(b) $x'' + (1 - x^2 - x'^2)x' + x = 0$

4. (*Hopf bifurcation*) (a) Show that the nonlinear system

$$x' = \epsilon x + y - x(x^2 + y^2), \tag{4.1a}$$

$$y' = -x + \epsilon y - y(x^2 + y^2) \tag{4.1b}$$

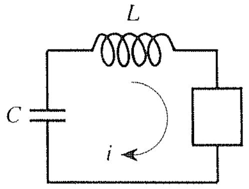
can be simplified to

$$r' = r(\epsilon - r^2), \tag{4.2a}$$

$$\theta' = -1 \tag{4.2b}$$

by the change of variables $x = r\cos\theta$, $y = r\sin\theta$ from the Cartesian x, y variables to the polar r, θ variables. HINT: Putting $x = r\cos\theta$, $y = r\sin\theta$ into (4.1) gives differential equations each of which contains both r' and θ' on the left-hand side. Suitable linear combinations of those equations give (4.2a,b), respectively. We suggest that you use the shorthand $\cos\theta \equiv c$ and $\sin\theta \equiv s$ for brevity.
(b) From (4.2) show that the origin in the x, y plane is a stable focus if $\epsilon < 0$ and an unstable focus if $\epsilon > 0$, and show that working from (4.1), instead, one obtains the same classification.
(c) Show from (4.2) that $r(t) = \sqrt{\epsilon}$ is a trajectory (if $\epsilon > 0$) and, in fact, a stable limit cycle. NOTE: Observe that zero is a bifurcation value of ϵ. As ϵ increases, a limit cycle is born as ϵ passes through the value zero, and its radius increases with ϵ. This is known as a **Hopf bifurcation**.
(d) Modify (4.1) so that it gives an *un*stable limit cycle at $r = \sqrt{\epsilon}$, instead.

5. The box in the circuit shown in the figure represents an "active element" such as a semiconductor or vacuum tube, the voltage drop across which is a known function $f(i)$ of the current i. Thus, Kirchhoff's voltage law gives $L\dfrac{di}{dt} + f(i) + \dfrac{1}{C}\displaystyle\int i\,dt = 0$.



(a) If f is of the form $f(i) = ai^3 - bi$, show that one obtains

$$Li'' + (3ai^2 - b)i' + \frac{1}{C}i = 0. \tag{5.1}$$

(b) Show that by a suitable scaling of both the independent and dependent variables one can obtain from (5.1) the van der Pol equation

$$I'' - \epsilon(1 - I^2)I' + I = 0, \tag{5.2}$$

where primes denote differentiation with respect to the new time variable τ, where $t = \alpha\tau$ and $i = \beta I$. That is, find α, β, and ϵ in terms of L, C, a, and b.

6. (*Rayleigh's equation and van der Pol relaxation oscillation*) (a) Show that if we set $x = z'$ in the van der Pol equation $x'' - \epsilon(1 - x^2)x' + x = 0$ and integrate once, we obtain $z'' - \epsilon(z' - z'^3/3) + z = C$, where C is a constant. Setting $z = u + C$, to absorb the C, obtain **Rayleigh's equation**

$$u'' - \epsilon\left(1 - \frac{u'^2}{3}\right)u' + u = 0 \tag{6.1}$$

on $u(t)$. The latter was studied by *Lord Rayleigh* (John William Strutt, 1842–1919) in connection with the vibration of a clarinet reed.
(b) Letting $u' = v$, reduce (6.1) to a system of two equations. Show that the only singular point of that system is $(0, 0)$, and classify its type.
(c) Choosing initial values for u and v, use computer software to obtain phase portraits and plots of $u(t)$ versus t, much as we have in Fig. 2 and 3, for $\epsilon = 0.2, 1$, and 5. For each of those ϵ's, estimate the amplitude and period of the limit cycle solution.
(d) To study the relaxation oscillation ($\epsilon \to \infty$) of the van der Pol equation it is more convenient to work with the Rayleigh equation (6.1), as we shall see. With $u' = v$, let us scale the u variable according to $u = \epsilon w$. Show that the resulting system is

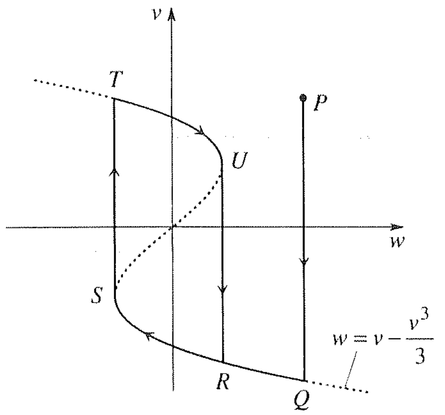
$$\epsilon w' = v, \tag{6.2a}$$

$$v' = \epsilon\left(v - \frac{v^3}{3}\right) - \epsilon w \tag{6.2b}$$

so

$$\frac{dv}{dw} = \epsilon^2 \frac{(v - v^3/3) - w}{v}. \tag{6.3}$$

As $\epsilon \to \infty$ we see from (6.3) that $dv/dw \to \infty$ at all points in the v, w phase plane except on the curve $v - v^3/3 - w$. So for any initial point, say P, explain why the solution is as shown in the figure. For instance, why is the direction downward from P, and why does the trajectory jump from S to T and from U to R? The loop $RSTUR$ is traversed repeatedly and is the limit cycle. Finally, use the figure to sketch $v(t)$ and hence the limit cycle solution $x(t)$ of the van der Pol equation (for $\epsilon \to \infty$). The result should be similar to the $\epsilon = 5$ part of Fig. 3. (HINT: Recall the expression "$s' = a$" for the phase velocity in Exercise 8 of Section 7.2.) Finally, explain why it is convenient to work with the Rayleigh equation in order to find the relaxation oscillation of the van der Pol equation.



7. In connection with Fig. 1b we suggested that over the limit cycle the energy gain, while the representative point is in the strip $|x| < 1$, exactly balances the energy loss while the point is outside of that strip. Let us explore that idea.

(a) Multiplying (1) through by dx and integrating over one cycle, show that

$$\frac{1}{2}\,x'^2\Big|_i^f - \epsilon\int_i^f (1-x^2)x'\,dx + \frac{1}{2}\,x^2\Big|_i^f = 0, \tag{7.1}$$

where i and f simply denote the initial point and final point, respectively. Explain, further, why (7.1) reduces to

$$\int_i^f (1-x^2)x'\,dx = 0, \tag{7.2}$$

which expresses the balance stated above – namely, that the net work done over one cycle by the $\epsilon(1-x^2)x'$ term in (1) is zero.

(b) For the case of small ϵ $(0 < \epsilon \ll 1)$, seek the limit cycle solution in the form $x(t) \approx a\cos t$. Putting the latter into (7.2), show that one obtains $a = 2$ as the radius of the circular limit cycle, as claimed in the text. NOTE: Put differently, (7.1) is equivalent to

$$E\Big|_i^f = \epsilon\int_i^f (1-x^2)x'\,dx, \tag{7.3}$$

where $E = \frac{1}{2}x'^2 + \frac{1}{2}x^2$. That is, the change in the total energy E over one cycle is equal to the net work done by the force $\epsilon(1-x^2)x'$. For a periodic motion $E\big|_i^f$ is zero which, once again, gives (7.2).

7.6 The Duffing Equation: Jumps and Chaos

7.6.1. Duffing equation and the jump phenomenon. Besides the van der Pol equation, also of great importance is the **Duffing equation**:

$$\boxed{mx'' + rx' + \alpha x + \beta x^3 = F_0\cos\Omega t,} \tag{1}$$

studied by *G. Duffing* around 1918. Whereas primary interest in the van der Pol equation is in the unforced equation and its self-excited limit cycle oscillation, most of the interest in the Duffing equation involves the various steady-state oscillations that can arise in response to a harmonic forcing function such as $F_0\cos\Omega t$.

Physically, (1) arises in modeling the motion of a damped, forced, mechanical oscillator of mass m having a nonlinear spring. That is, the spring force is not kx but $\alpha x + \beta x^3$. We assume that $\alpha > 0$ but that β can be positive (for a "hard spring") or negative (for a "soft spring").

The linear version of (1), where $\beta = 0$, is discussed in Section 3.8, and an important result there consisted of the amplitude response curves – namely, the graphs of the amplitude of the steady-state vibration versus the driving frequency Ω for various values of F_0. For the linear case we obtained two linearly independent homogeneous solutions and a particular solution and used linearity and superposition to form from them the general solution, which contained all possible solutions. Understand that because equation (1) is nonlinear, that approach is not applicable.

Consider first the undamped case ($r = 0$), and let $m = 1$ for simplicity, so (1) becomes

$$x'' + \alpha x + \beta x^3 = F_0 \cos \Omega t. \tag{2}$$

Further, suppose that β is small. Since (1) is nonlinear, we expect it to have a wealth of different sorts of solutions. Of these, particular interest attaches to the so-called **harmonic response**

$$x(t) \approx A \cos \Omega t \qquad \text{(for } \beta \text{ small)} \tag{3}$$

at the same frequency as the driving force. As shown in the exercises, pursuing an approximate solution of the form (3) yields the amplitude-frequency relation

$$\Omega^2 = \alpha + \frac{3}{4}\beta A^2 - \frac{F_0}{A}, \tag{4}$$

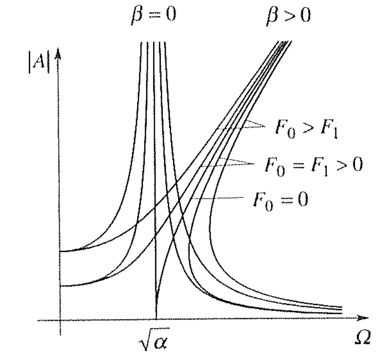


Figure 1. Amplitude response curves; undamped.

which gives the response curves shown in Fig. 1. For the linear case, where $\beta = 0$, (4) reduces to $A = F_0/(\alpha - \Omega^2)$, which result agrees with the amplitude-frequency relation found in Section 3.8. For $\beta > 0$ the curves bend to the right (shown), and for $\beta < 0$ they bend to the left (not shown). Thus, the effect of the nonlinear βx^3 term in (2) is to cause the response curves to bend to one side or the other. Recall from Section 3.8 that for the linear system ($\beta = 0$), the amplitude $|A|$ is infinite when the system is driven at its natural frequency $\sqrt{\alpha}$. [More precisely, we saw that the solution form $x(t) = A \cos \Omega t$ simply does not work for the system $x'' + \alpha x = F_0 \cos \Omega t$ if $\Omega = \sqrt{\alpha}$, but that the method of undetermined coefficients gives $x(t) = \dfrac{F_0}{2\Omega} t \sin \Omega t$, which growing oscillation is known as resonance.] However, because of the bending of the response curves, resonance cannot occur in the nonlinear case. That is, we see from Fig. 1 that if $\beta \neq 0$ then at each driving frequency Ω the response amplitude $|A|$ is finite.

What is the effect of including damping, of letting r be positive in (1) rather than zero? In that case we need to allow for a phase shift (as in Section 3.8), and seek $x(t) \approx A \cos(\Omega t + \Phi)$ in place of (3). The result would be a modified amplitude- frequency relation, in place of (4), and a "capping off" of the response curves as shown in Fig. 2a.

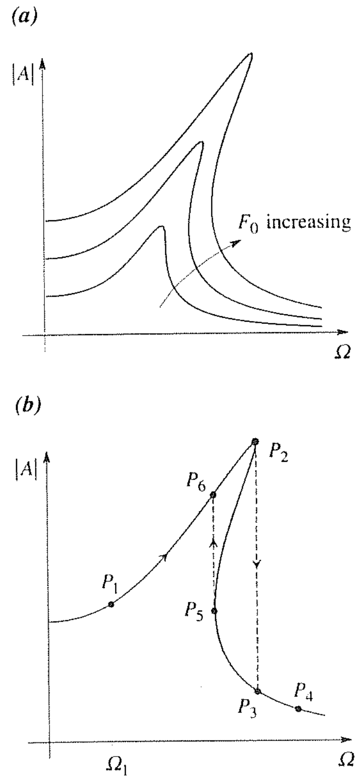


Figure 2. Amplitude response curves; damped.

If $\Omega = \Omega_1$, for instance, then the response is at P_1 in Fig. 2b. Suppose we can vary the driving frequency Ω continuously by turning a control knob, like the volume knob on a radio. If we increase Ω slowly (remember that Ω is regarded as a constant in this analysis, so we need to increase it *very* slowly), then the representative point moves to the right along the response curve. But what happens when it reaches P_2, where the response curve has a vertical tangent? Numerical simulation reveals that the point jumps down to P_3, where it can then continue moving rightward on the response curve if Ω is increased further. That is, there is a transient period during which the oscillation changes from the large amplitude at P_2 to the small amplitude at P_3. Suppose that once the oscillation has settled down to the new amplitude we increase Ω further and stop at P_4. If we now *de*crease Ω the representative point does *not* jump up from P_3 to P_2. Rather, it continues to the point of vertical tangency at P_5, then jumps up to P_6 and continues, say to P_1. Thus, we meet the **jump phenomenon** associated with the Duffing equation, whereby a continuous variation of a system parameter (Ω) can lead to a discontinuous change in the output, namely, the jumps in amplitude from P_2 to P_3 and from P_5 to P_6.

Observe that the middle branch of the response curve, between P_5 and P_2, is inaccessible! That is, if we vary Ω so that the representative point moves from P_1 to P_4 and back again, it never moves along the middle branch.

What if we start the system at an Ω between the vertical dashed lines? Which of the three possible amplitudes will result? It can be shown that the solutions to (1) corresponding to points on the middle branch (P_5P_2) are unstable in essentially the same way that a hilltop is an unstable equilibrium point for a marble, while solutions corresponding to points on the upper and lower branches are stable in essentially the same way that the bottom of a valley is a stable equilibrium point for the marble. Thus, solutions corresponding to points on the middle branch will not be observed – either in the physical system or in its mathematical simulation. If we start the system at an Ω between the vertical dashed lines, and keep Ω fixed, the steady-state oscillation achieved will correspond to a point on the lower branch or on the upper branch. Given the initial conditions, it is possible to predict which of the two branches will "attract" the motion, but that story is well beyond our present scope.*

Both the limit cycle phenomenon exhibited by the van der Pol equation and the jump phenomenon exhibited by the Duffing equation are possible only for nonlinear equations; neither phenomenon could occur for a linear differential equation.

The jump from P_2 to P_3 in Fig. 2b is reminiscent of a Stock Market crash, such as occurred in 1929, yet there is no analog, in Stock Market behavior, of the *up*ward jump from P_5 to P_6. However, observe that in Fig. 2b there is only one control parameter, Ω. If there were two (or more), say Ω and Ψ, then instead of the curve in Fig. 2b we would have an $|A|$ surface above an Ω, Ψ plane. In that case it might be possible, by suitable variation of Ω and Ψ in the Ω, Ψ "control plane," to climb the mountain in a continuous manner, jump off the cliff at P_2, climb the mountain again, and so on, so that only downward jumps were obtained.

Finally, recall from Section 7.5 that the central nervous system is highly nonlinear. Thus, we might well be on the lookout for Duffing-like jumps in human (or animal) behavior due to continuous variation of one or more control parameters. Such jumps are indeed observed! For instance, if one bullies a dog into a corner it will retreat, but at some point it may "turn" on its tormentor, with its behavior jumping, instantaneously, from retreating to attacking. Similarly with human behavior. Two more illustrations: First, observe that our psychological moods likewise tend to change rather abruptly – compared to how long they last. Second, note that

*See Section 7.2 of D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, 2nd ed. (Oxford: Oxford University Press, 1987).

someone with an eating disorder might fast for a period of time, then jump almost instantaneously to binging, and vice versa. For a readable account of the modeling of such systems as these, see E. C. Zeeman's *Catastrophe Theory* (Reading, MA: Addison-Wesley, 1977).

7.6.2. Chaos. Consider the physical system shown in Fig. 3, a box containing a slender vertical steel reed cantilevered from the "ceiling" and two magnets attached to the "floor"; x is the horizontal deflection of the end of the reed. In equilibrium, the reed is stationary, with its tip at one magnet or the other. However, if we vibrate the box periodically in the horizontal direction, we can (if the magnets are not too strong) shake the reed loose from its equilibrium position and set it into motion. This experiment was carried out by F. Moon and P. Holmes,* to study chaos. They modeled the system by the Duffing equation

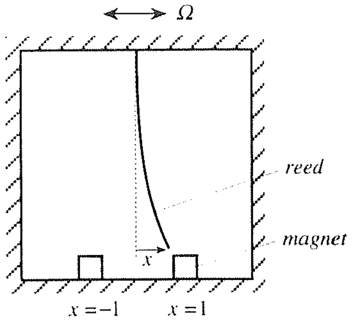


Figure 3. The "Moon beam" system of Moon and Holmes.

$$x'' + rx' - x + x^3 = F_0 \cos \Omega t, \tag{5}$$

where r is a (positive) damping coefficient and F_0 and Ω are the forcing function strength and frequency, respectively. The $-x + x^3$ terms approximate the force induced on the reed by the two competing magnets. It is zero at $x = 0, \pm 1$, which are therefore equilibrium points. To classify the equilibrium points, consider the unforced equation $x'' + rx' - x + x^3 = 0$ or, equivalently,

$$x' = y, \tag{6a}$$
$$y' = -ry + x - x^3. \tag{6b}$$

We leave it for the exercises for you to show that the origin $(x, y) = (0, 0)$ is an unstable equilibrium point, namely a saddle, and that $(\pm 1, 0)$ are stable equilibrium points (stable foci if $r < \sqrt{8}$ and stable nodes if $r > \sqrt{8}$). For the undriven system ($F_0 = 0$), imagine displacing the reed tip to $x = 0.5$, say, and releasing it from rest. Then it will undergo a damped motion about $x = 1$. If instead we release it from $x = -0.5$, say, it will undergo a damped motion about $x = -1$.

What will happen if we *force* the system ($F_0 > 0$)? We can imagine the reed undergoing a steady-state oscillation about $x = +1$ or $x = -1$, depending upon the initial conditions. To encourage physical insight, it is useful to consider the potential energy $V(x)$ associated with the magnetic force $F(x) = x - x^3$. Recalling from Section 7.4.2 that $F(x) = -V'(x)$, we have

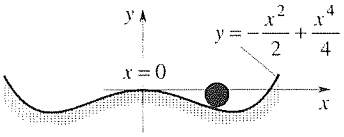


Figure 4. Double well.

$$V(x) = -\frac{x^2}{2} + \frac{x^4}{4}. \tag{7}$$

Thus, in place of a reed/magnet system we can conceptualize the system more intuitively as a mass in a double well as sketched in Fig. 4. That is, its gravitational potential energy is $V(x) = mgy = mg(-x^2/2 + x^4/4)$ which, except for the scale factor mg, is the same as $V(x)$ for the reed/magnet system, given in (7).

*F. C. Moon and P. J. Holmes, A Magnetoelastic Strange Attractor, *Journal of Sound and Vibration*, Vol. 65, pp. 275–296.

If the mass sits at the bottom of a well, then if we vibrate the system horizontally we expect the mass to oscillate within that well. If we vibrate so energetically that the mass jumps from one well to the other, then the situation becomes more complicated.

Rather than carry out the physical experiment, let us simulate it numerically by using computer software to solve (5). Let us fix $r = 0.3$ and $\Omega = 1.2$, and use the initial conditions $x(0) = 1$ and $x'(0) = 0.2$, say, so that if F_0 is not too large we expect an oscillation near $x = 1$ (i.e., in the right-hand well).* Our plan is to use successively larger F_0's and see what happens. The results are quite striking (like those obtained by Moon and Spencer). They are presented in Fig. 5, both in the x, y phase plane and as $x(t)$ versus t.

In Fig. 5a we set $F_0 = 0.2$ and find that after a brief transient period there results a steady-state oscillation near $x = 1$, as anticipated. That oscillation is of the same period as the forcing function (namely, $2\pi/\Omega = 2\pi/1.2 \approx 5.236$) so it is called a **period-1 oscillation**, or **harmonic oscillation**. Period-1 oscillations persist up to around $F_0 = 0.27$, but for $F_0 > 0.27$ different solution types arise.

For $F_0 = 0.28$ the solution is still periodic, but it takes two loops (in the x, y plane) to complete one period and the period is now doubled, namely, $4\pi/\Omega$. Thus, it is called a **period-2 oscillation**, or a **subharmonic oscillation of order 1/2**. In Fig. 5b–5f we omit the transient and display only the steady-state periodic solution so as not to obscure that display. Observe, in Fig. 5b, the point where the trajectory crosses itself. That crossing does not violate the existence and uniqueness theorem given in Section 7.3.1 because equation (5) is nonautonomous and the phase plane figure shows the *projection* of the non-self-intersecting curve in three-dimensional x, y, t space onto the x, y plane.

If we increase F_0 further, to 0.29, the forcing is sufficiently strong so that during the transient phase the mass (reed) is driven out of the right-hand potential well and ends up in a period-4 oscillation about the left-hand well. To observe this result we need to be patient and run the calculation to a sufficiently large time, namely, beyond $t \approx 400$. This period doubling continues as F_0 increases from 0.29 up to around 0.30. For $F_0 > 0.30$ a period-5 oscillation results that now encompasses both stable equilibrium points (Fig. 5d).

The regime $0.37 < F_0 < 0.65$ is found to be rather chaotic, with essentially random motions, as seen in Fig. 5e for the case $F_0 = 0.5$.

Reviewing these results, observe that as we increased F_0 the period of the motion increased until, when F_0 was increased above 0.37, periodicity was lost altogether and the motion became chaotic. (We can think of that motion as periodic but with infinite period.) It would be natural to expect that a further increase of F_0 would lead to even greater chaos (if one were to quantify degree of chaos), yet we find that if F_0 is increased to 0.65 we once again obtain a periodic solution, namely, the period-2 solution shown in Fig. 5f, and if F_0 is increased further to 0.73 then we obtain a period-1 solution (not shown).

In summary, we see that the forced Duffing equation admits a great variety

*These parameter values are the same as those chosen in Section 12.6 of D. W. Jordan and P. Smith (ibid). We refer you to that source for a more detailed discussion than we offer here.

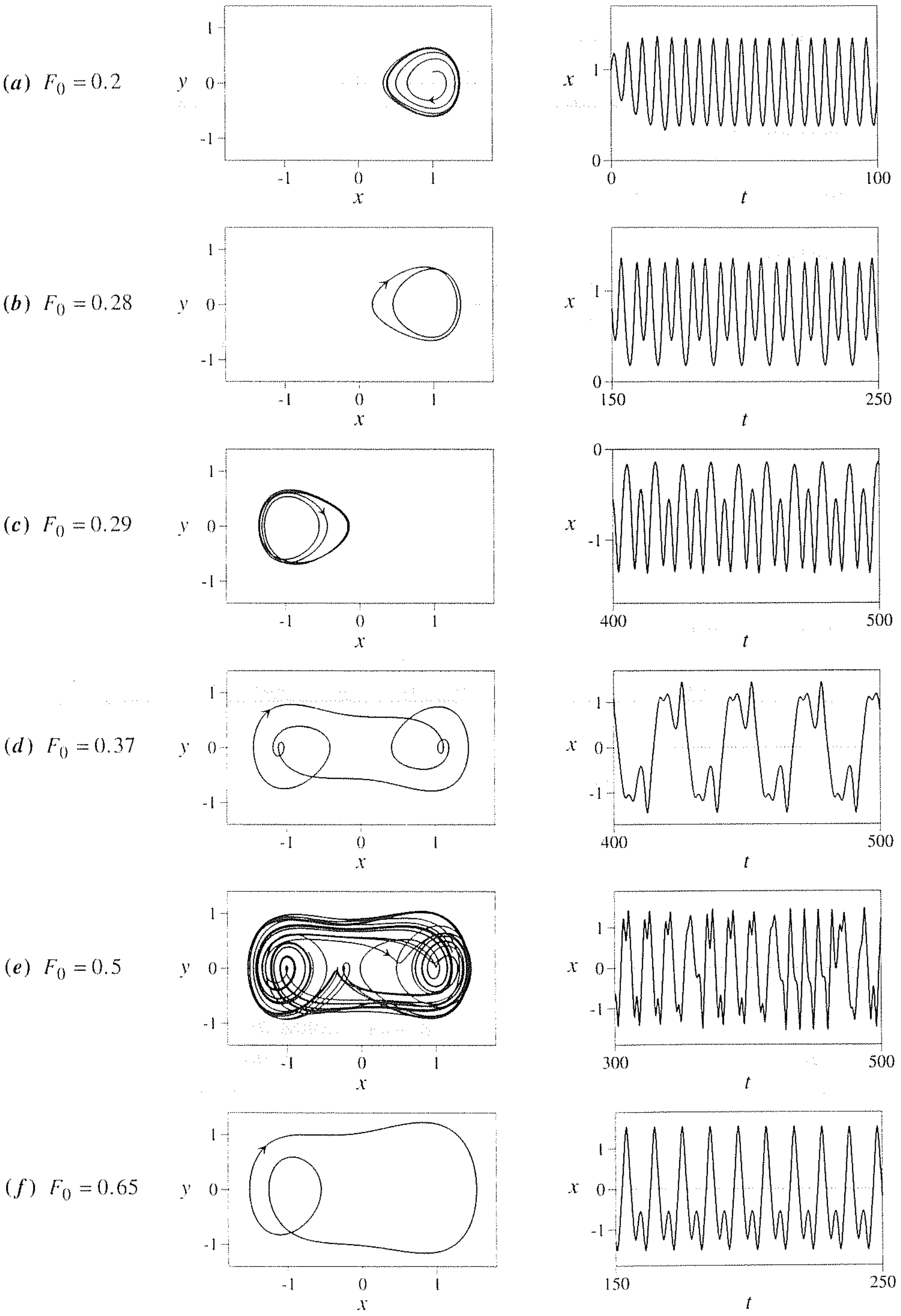


Figure 5. Response to (5): $r = 0.23$, $\Omega = 1.2$.

of periodic solutions and chaotic ones as well, and that these different regimes correspond to different intervals on an F_0 axis. (We chose to hold r and Ω fixed and to vary only F_0, but we could have varied r and/or Ω as well.) It is possible to predict analytically how the solution type varies with F_0, r, and Ω, but that analysis is beyond the scope of this introductory discussion.*

Having classified the response in Fig. 5e as chaotic, it behooves us to clarify what we mean by that. A reasonable working definition of **chaos** is *behavior in a deterministic system, over time, which is aperiodic and which is sensitive to initial conditions.*

By a system being **deterministic** we mean that the governing differential equation(s) and initial conditions imply the existence of a unique solution over subsequent time. Whether we are able to find that solution analytically, or whether we need to use computer simulation. is immaterial. For instance, for given values of r, F_0, Ω, the system consisting of equation (5), together with initial conditions x(0) and $x'(0)$, is deterministic. The choice $r = 0.3$, $F_0 = 0.37$, $\Omega = 1.2$, $x(0) = 1$, and $x'(0) = 0.2$, say, implies the unique response shown in Fig. 5d. If we rerun the numerical solution or solve the problem analytically (if we could), we obtain the same solution as shown in the figure. Likewise even for the chaotic response shown in Fig. 5e.

By the response being **aperiodic** we simply mean that it does not approach a periodic solution or a constant.

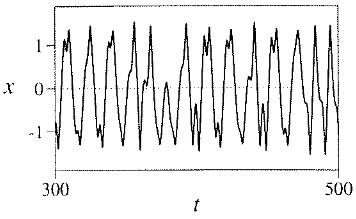


Figure 6. Sensitivity to initial conditions.

To illustrate what is meant by *sensitivity to initial conditions*, let us rerun the case corresponding to Fig. 5e, but with the initial conditions changed from $x(0) = 1$ and $x'(0) = 0.2$ to $x(0) = 1$ and $x'(0) = 0.2000000001$. Observe that the results (Fig. 6) bear virtually no resemblance to those in Fig. 5e. This circumstance is of great significance because if the initial conditions are known to only six significant figures, say, then the task of predicting the response is hopeless!

Another well known example of chaos is provided by the **Lorenz equations**

$$\begin{aligned} x' &= p(y - x), \\ y' &= (q - z)x - y, \\ z' &= xy - rz, \end{aligned} \tag{8}$$

where p, q, r are constants. This system was studied by the mathematical meteorologist *E. Lorenz* in 1963 in connection with the Bénard problem, whereby one seeks the effect of heating a horizontal fluid layer from below.† That problem is of fundamental interest in meteorology because the earth, having been heated by the sun during the day, radiates heat upward into the atmosphere in the evening, thus destabilizing the atmospheric layer above it. Lorenz's contribution was in discovering the chaotic nature of the solution for certain ranges of the physical parameters p, q, r, thereby suggesting the *impossibility* of meaningful long-range weather prediction. Some discussion of (8) is left for the exercises.

*See Section 12.6 in Jordan and Smith (ibid).

†E. Lorenz, Deterministic Nonperiodic Flows, *Journal of Atmospheric Sciences*, Vol. 20, pp. 130–141, 1963.

Perhaps the classic problem of chaos is that of turbulence, in fluid mechanics, be it in connection with the chaotic eddies and mixing behind a truck on a highway or the turbulent breakup of a rising filament of smoke.

To appreciate the revolution in physics that has resulted from recent work on chaos, one needs to understand the euphoria that greeted the birth of Newtonian mechanics and the calculus, according to which both the past and the future are contained in the system of differential equations and initial conditions at the present instant. In the words of Ivar Ekeland,* "Past and future are seen as equivalent, since both can be read from the present. Mathematics travels back in time as easily as a wanderer walks up a frozen river." As Ekeland points out, that statement is not quite true for, as we now know and as was understood by Poincaré even a century ago, deterministic nonlinear systems can turn out to be chaotic, in which case they are useless for long-term prediction.

Closure. The common thread in this section is the Duffing equation (1). For $\alpha > 0$ we study (1) in connection with the reed/magnet system of Moon and Holmes, shown in Fig. 3. Numerical solution of the governing equation (5), for a sequence of increasing F_0 values, leads to a variety of solution types: a harmonic response, various subharmonic responses, and even chaotic responses. The approach to chaos, as we increase F_0, is typical in that the onset of chaos is preceded by a sequence of period doublings.

We define chaos as behavior in a deterministic system (which must be nonlinear if it is to exhibit chaos) over time, which is aperiodic and so sensitive to initial conditions that accurate long-range predictions of the solution are not possible.

Computer software. To generate the responses shown in Fig. 5 and 6 we use the *Maple* phaseportrait command. However, it is worth mentioning how we obtain the graphs in Fig. 1 and 2, because (4) does not give A explicitly but only *im*plicitly as a function of Ω. For instance, suppose we wish to plot the graphs of y versus x, over $0 < x < 3$ and $0 < y < 4$, for the functions $y(x)$ given implicitly by the equations $y - y^3 = x$ and $4y - y^3 = x$. First, enter

```
with(plots):
```

and return, to access the subsequent plotting command. Then use the **implicitplot** command. Enter

```
implicitplot({y - y^3 = x, 4 * y - y^3 = x}, x = 0..3, y = 0..4,
numpoints = 1000);
```

and return. Here, numpoints indicates the number of points to be used. To plot the single function $y - y^3 = x$, use $y - y\hat{}3 = x$ in place of implicitplot($\{y - y\hat{}3 = x,\ 4 * y - y\hat{}3 = x\}$.

[*] *Mathematics and the Unexpected* (Chicago: Chicago University Press, 1988).

EXERCISES 7.6

1. *(Derivation of the Duffing amplitude-frequency relation)* We state in the text that if one seeks an approximate harmonic solution $x(t) \approx A\cos\Omega t$ to

$$x'' + \alpha x + \beta x^3 = F_0 \cos\Omega t, \tag{1.1}$$

then one obtains the ampitude-frequency relation

$$\Omega^2 = \alpha + \frac{3}{4}\beta A^2 - \frac{F_0}{A} \tag{1.2}$$

discovered by Duffing. A modern derivation of (1.2) would probably use a so-called *singular perturbation method* of *strained variables*, but here we will pursue a simpler iterative approach which is essentially that of Duffing; namely, we replace (1.1) by the iterative scheme

$$x''_{n+1} = -\alpha x_n - \beta x_n^3 + F_0 \cos\Omega t, \tag{1.3}$$

choose the initial iterate as $x(t) = A\cos\Omega t$, and then use (1.3) to find the successive iterates $x_1(t), x_2(t), \dots$. It is surely not obvious whether that procedure will work, so it makes sense to try it out first for the simple *linear* case where $\beta = 0$, for which we know the exact solution.

(a) In that case ($\beta = 0$), show that if we seek a harmonic solution $x(t) = A\cos\Omega t$ of the Duffing equation (1.1) with $\beta = 0$, we obtain $A = F_0/(\alpha - \Omega^2)$, and hence the exact solution

$$x(t) = \frac{F_0}{\alpha\Omega^2}\cos\Omega t. \tag{1.4}$$

(b) Next, use (1.3) to generate $x_1(t), x_2(t), \dots$, for $\beta = 0$, and show that

$$\begin{aligned} x_1(t) &= \frac{\alpha A - F_0}{\Omega^2}\cos\Omega t, \\ &\vdots \\ x_n(t) &= \left\{\left(\frac{\alpha}{\Omega^2}\right)^n A \right. \\ &\left. -\frac{F_0}{\Omega^2}\left[1 + \left(\frac{\alpha}{\Omega^2}\right) + \cdots + \left(\frac{\alpha}{\Omega^2}\right)^{n-1}\right]\right\}\cos\Omega t. \end{aligned} \tag{1.5}$$

That is, put $x_0(t) = A\cos\Omega t$ into the right side of (1.3) and integrate twice to obtain x_1. Then put that x_1 into the right side of (1.3) and integrate twice to obtain x_2, and so on. By the time you reach x_3, the general result shown in (1.5) should be apparent.

(c) Recalling that the geometric series $1 + x + x^2 + x^3 + \cdots$ converges to $1/(1-x)$ if $|x| < 1$ and diverges otherwise, show that the $x_n(t)$ sequence given by (1.5) does indeed converge to the exact solution (1.4) as $n \to \infty$, provided that $|\alpha/\Omega^2| < 1$.

(d) In fact, show that if we equate the coefficients of $\cos\Omega t$ in $x_0(t) = A\cos\Omega t$ and $x_1(t) = \dfrac{\alpha A - F_0}{\Omega^2}\cos\Omega t$, we happen to obtain $A = F_0/(\alpha - \Omega^2)$, which agrees with the exact solution (1.4)!

(e) In view of the striking success in part (d), we are encouraged to expect good results even for the nonlinear case where $\beta \neq 0$. Thus, put $x_0(t) = A\cos\Omega t$ into the right side of (1.3) and integrate twice to obtain $x_1(t)$. Then, as in (d), equate the coefficients of $\cos\Omega t$ in $x_1(t)$ and $x_0(t)$, and show that the result is Duffing's relation (1.2). HINT: The identity $\cos^3\theta = (3\cos\theta + \cos 3\theta)/4$ should be helpful.

2. *(Computer problem regarding the Duffing jump phenomenon)* For the undamped case the amplitude-frequency relation is given by (4). For the damped case it is given by

$$F_0^2 = \left[\left(\alpha - \Omega^2\right)A + \frac{3}{4}\beta A^3\right]^2 + (r\Omega A)^2. \tag{2.1}$$

Throughout parts (a)–(g) let $F_0 = 2$, $\alpha = 1$, $\beta = 0.4$, and $r = 0.3$, for definiteness.

(a) Use (2.1) to generate a computer plot of $|A|$ versus Ω as we did in Fig. 2b. NOTE: Actually, there is no need to distinguish between A and $|A|$ since, unlike (4), (2.1) contains only even powers of A.

(b) For $\Omega = 1$ solve (2.1) for A. (HINT: Using *Maple*, for instance, use the fsolve command.) Next, use computer software such as the *Maple* phaseportrait command to solve

$$x'' + rx' + \alpha x + \beta x^3 = F_0 \cos\Omega t, \tag{2.2}$$

and plot $x(t)$ versus t over a sufficiently long time interval to obtain the steady-state response. Compare the amplitude of the resulting steady-state response with the value of A obtained from (2.1).

(c) Same as (b) but for an Ω point specified by your instructor, to the left of the first point of vertical tangency ($\Omega \approx 1.71$).

(d) Same as (b), but for an Ω point specified by your instructor, to the right of the second point of vertical tangency ($\Omega \approx 2.05$).

(e) Now consider an Ω between the two points of vertical tangency, say, $\Omega = 1.8$. Solve (2.1) for the three A values. Next, use computer software to solve (2.2) over a sufficiently long

time interval to obtain the steady-state response. Depending upon the initial conditions that you impose, you should obtain the smallest or largest of the three A values, but never the middle one. Keeping $x'(0) = 0$, determine the approximate value of $x(0)$ below which you obtain the small-amplitude response and above which you obtain the large-amplitude response.
(f) Same as (e) but for an Ω point specified by your instructor.
(g) Continuing to use the r, α, β, F_0 values given above, now let Ω be slowly varying according to $\Omega = 1.9 - 0.0005t$, and solve (2.2) over $0 < t < 800$ with the initial conditions $x(0) = x'(0) = 0$. Plot the resulting $x(t)$ versus t and discuss your results in terms of the ideas discussed in this section.

3. (a) Refer to (4) and Fig. 1. What is the asymptotic form of the graph of $|A|$ versus Ω as $|A| \to \infty$?
(b) In Fig. 1 we show several amplitude response curves for $\beta = 0$ and $\beta > 0$ and for several values of F_0. Obtain the analogous curves for the case where $\beta < 0$, either by a careful hand sketch or by computer plotting.

4. Determine the location and type of any singular points of (6).

5. (a)–(f) Obtain x, y and x, t plots for the cases depicted in Fig. 5. Your results should be the same as those in Fig. 5.

6. We stated that "period doubling continues as F_0 increases from 0.29 up to around 0.30." Find, by numerical experimentation, an F_0 that gives a period-8 oscillation (and, if possible, period-16) and obtain computer plots analogous to those in Fig. 5.

7. We stated that "if F_0 is increased further to 0.73, then we obtain a period-1 solution." Obtain computer plots for that case, analogous to those in Fig. 5.

8. We found an extreme sensitivity to initial conditions for the chaotic regime. Specifically, the plot in Fig. 6 bears little resemblance to the corresponding one in Fig. 5e, even though the initial conditions differed by only 10^{-10}. Show that that sensitivity is *not* found for the non-chaotic responses – namely, for the periodic responses. Specifically, rerun the cases reported in Fig. 5d and 5f, but with $x'(0) = 0.2$ changed to 0.20001, say. Do your results appear to reproduce those in Fig. 5?

9. The equation

$$x'' + 0.3x' + \sin x = F_0 \cos t \tag{9.1}$$

is similar to the one occurring in the Moon/Holmes experiment.

(a) Describe a physical problem that would have a governing equation of motion of that form. (We have assigned numerical values to all of the physical parameters except to F_0, which we leave for the purpose of numerical experimentation.)
(b) We leave this problem a bit more open ended than the foregoing ones, and simply ask you to carry out an analytical and experimental study of (9.1). For instance, you might investigate the singular points of the homogeneous version of (9.1), and also run computer solutions for a range of F_0 values, somewhat as done for equation (5).

Chapter 7 Review

In Sections 7.2–7.5 we study the autonomous system

$$x' = P(x,y), \qquad y' = Q(x,y) \tag{1}$$

mostly in the x, y phase plane. We focus considerable attention on the singular points of (1), the points (if any) where $P(x,y) = 0$ and $Q(x,y) = 0$. Linearizing equations (1) about each singular point, we obtain a simpler system of the form

$$\begin{aligned} X' &= aX + bY \\ Y' &= cX + dY, \end{aligned}$$

where $X = x - x_s$, $Y = y - y_s$, and (x_s, y_s) is the singular point. Considering only elementary singular points (that is, for which $ab - cd \neq 0$), we classify them as centers, foci, nodes, and saddles. The Hartman–Grobman theorem assumes that the linearized system faithfully captures the nature of the local flow (except possibly for the borderline cases of proper nodes and centers, as explained in Section 7.4.1).

In Section 7.4 we study applications and introduced the idea of a bifurcation, whereby the behavior of the system changes qualitatively as a system parameter passes through a critical value. We illustrate the concept with an example of a saddle-node bifurcation from molecular biology.

In Section 7.5 we study the van der Pol equation

$$x'' - \epsilon(1 - x^2)x' + x = 0, \qquad (\epsilon > 0)$$

which introduce us to the concept of limit cycles and relaxation oscillations.

Finally, we study the forced Duffing equation

$$mx'' + rx' + \alpha x + \beta x^3 = F_0 \cos \Omega t$$

in two contexts. First, we consider it as modeling a mechanical oscillator, with nonlinear spring force $\alpha x + \beta x^3$, where $\alpha > 0$. Of the various possible solutions that can be obtained from different initial conditions, we study only the harmonic response – that is, the steady-state periodic response at the same frequency as the driving frequency Ω. The key feature that was revealed was the bending of the amplitude response curves and the resulting jump phenomenon, whereby the response amplitude jumps as Ω is increases or decreases slowly through a critical value.

We also consider it as modeling the "double well" reed/magnet system of Moon and Holmes. By numerical simulation, we find that if F_0 is not too large, then the oscillation is confined to one of the two wells. As F_0 is increased, there results a sequence of period doublings, giving so-called subharmonic responses, until F_0 becomes large enough to drive the response out of that well. Beyond a critical F_0 value, we then obtain a chaotic response involving both wells.

Chapter 8

Systems of Linear Algebraic Equations; Gauss Elimination

8.1 Introduction

There are many applications in science and engineering where application of the relevant physical law(s) immediately produces a set of linear algebraic equations. For instance, the application of Kirchoff's laws to a DC electrical circuit containing any number of resistors, batteries, and current loops immediately produces such a set of equations on the unknown currents. In other cases, the problem is stated in some other form such as one or more ordinary or partial differential equations, but the solution method eventually leads us to a system of linear algebraic equations. For instance, to find a particular solution to the differential equation

$$y''' - y'' = 3x^2 + 5\sin x \tag{1}$$

by the method of undetermined coefficients (Section 3.7.2), we seek it in the form

$$y_p(x) = Ax^4 + Bx^3 + Cx^2 + D\sin x + E\cos x. \tag{2}$$

Putting (2) into (1) and equating coefficients of like terms on both sides of the equation gives five linear algebraic equations on the unknown coefficients $A, B, \ldots, E$. Or, solving the so-called Laplace partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{3}$$

on the rectangle $0 < x < 1, 0 < y < 1$ by the method of finite differences (which is studied in Section 20.5), using a mesh size $\Delta x = \Delta y = 0.05$, gives $19^2 = 361$ linear algebraic equations on the unknown values of u at the 361 nodal points of the mesh. Our point here is not to get ahead of ourselves by plunging into partial differential equations, but to say that the solution of practical problems of interest in science and engineering *often* leads us to systems of linear algebraic equations.

Such systems often involve a great many unknowns. Thus, the question of existence (Does a solution exist?), which often sounds "too theoretical" to the practicing engineer, takes on great practical importance because a considerable computational effort is at stake.

The subject of linear algebra and matrices encompasses a great deal more than the theory and solution of systems of linear algebraic equations, but the latter is indeed a central topic and is foundational to others. Thus, we begin this sequence of five chapters (8–12) on **linear algebra** with an introduction to the theory of systems of linear algebraic equations, and their solution by the method of Gauss elimination. Results obtained here are used, and built upon, in Chapters 9–12.

Chapters 9 and 10 take us from vectors in 3-space to vectors in n-space and generalized vector space, to matrices and determinants. Linear systems of algebraic equations are considered again, in the second half of Chapter 10, in terms of rank, inverse matrix, LU decomposition, Cramer's rule, and linear transformation. Chapter 11 introduces the eigenvalue problem, diagonalization, and quadratic forms; areas of application include systems of ordinary differential equations, vibration theory, chemical kinetics, and buckling. Chapter 12 is optional and brief and provides an extension of results in Chapters 9–11 to complex spaces.

8.2 Preliminary Ideas and Geometrical Approach

The problem of finding solutions of equations of the form

$$f(x) = 0 \tag{1}$$

occupies a place of both practical and historical importance. Equation (1) is said to be **algebraic**, or **polynomial**, if $f(x)$ is expressible in the form $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, where $a_n \neq 0$ for definiteness [i.e., if $f(x)$ is a polynomial of finite degree n], and it is said to be **transcendental** otherwise.

EXAMPLE 1. The equations $6x - 5 = 0$ and $3x^4 - x^3 + 2x + 1 = 0$ are algebraic, whereas $x^3 + 2\sin x = 0$ and $e^x - 3 = 0$ are transcendental since $\sin x$ and e^x cannot be expressed as polynomials of finite degree. ▮

Besides the algebraic versus transcendental distinction, we classify (1) as **linear** if $f(x)$ is a first-degree polynomial,

$$a_1 x + a_0 = 0, \tag{2}$$

and **nonlinear** otherwise. Thus, the first equation in Example 1 is linear, and the other three are nonlinear.

While (1) is one equation in one unknown, we often encounter problems involving more than one equation and/or more than one unknown – that is, a **system** of equations consisting of m equations in n unknowns, where $m \geq 1$ and $n \geq 1$,

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0, \\ f_2(x_1, \ldots, x_n) &= 0, \\ &\vdots \\ f_m(x_1, \ldots, x_n) &= 0 \end{aligned} \tag{3}$$

such as

$$\begin{aligned} x_1 - \sin(x_1 + 7x_2) &= 0, \\ x_2^3 + x_2 - 5x_1 + 6 &= 0. \end{aligned} \tag{4}$$

In (4) it happens that $m = n$ (namely, $m = n = 2$) so that there are as many equations as unknowns. In general, however, m may be less than, equal to, or greater than n so we allow for $m \neq n$ in this discussion even though $m = n$ is the most important case.

In this chapter we consider only the case where (3) is **linear**, of the form

$$\boxed{\begin{array}{lr} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = c_1, & \text{(eq.1)} \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = c_2, & \text{(eq.2)} \\ \vdots & \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = c_m, & \text{(eq.m)} \end{array}} \tag{5}$$

and we restrict m and n to be finite, and the a_{ij}'s and c_j's to be real numbers. If all the c_j's are zero then (5) is **homogeneous**; if they are not all zero then (5) is **nonhomogeneous**.

The subscript notation adopted in (5) is not essential but is helpful in holding the nomenclature to a minimum, in rendering inherent patterns more visible, and in permitting a natural transition to matrix notation. The first subscript in a_{ij} indicates the equation, and the second indicates the x_j variable that it multiplies. For instance, a_{21} appears in the second equation and multiplies the x_1 variable. To avoid ambiguity we should write $a_{2,1}$ rather than a_{21} so that one does not mistakenly read the subscripts as twenty-one, but we will omit commas except when such ambiguity is not easily resolved from the context.

We say that a sequence of numbers $s_1, s_2, \ldots, s_n$ is a **solution** of (5) if and only if each of the m equations is satisfied numerically when we substitute s_1 for x_1, s_2 for x_2, and so on. If there exist *one or more* solutions to (5), we say that the system is **consistent**; if there is precisely *one* solution, that solution is **unique**; and if there is more than one, the solution is **nonunique**. If, on the other hand, there are *no* solutions to (5), the system is said to be **inconsistent**. The collection of all solutions to (5) is called its **solution set** so, by "solving (5)" we mean finding its solution set.

Let us begin with the simple case, where $m = n = 1$:

$$a_{11}x_1 = c_1. \tag{6}$$

In the generic case, $a_{11} \neq 0$ and (6) admits the unique solution $x_1 = c_1/a_{11}$, but if $a_{11} = 0$ there are two possibilities: if $c_1 \neq 0$ then there are no values of x_1 such that $0x_1 = c_1$ and (6) is inconsistent, but if $c_1 = 0$ then (6) becomes $0x_1 = 0$, and $x_1 = \alpha$ is a solution for *any* value of α; that is, the solution is nonunique.

Far from being too simple to be of interest, the case where $m = n = 1$ establishes a pattern that will hold in general, for any values of m and n. Specifically, the system (5) will admit a unique solution, no solution, or an infinity of solutions. For instance, it will never admit 4 solutions, 12 solutions, or 137 solutions.

Next, consider the case where $m = n = 2$:

$$a_{11}x_1 + a_{12}x_2 = c_1, \qquad \text{(eq.1)} \tag{7a}$$
$$a_{21}x_1 + a_{22}x_2 = c_2. \qquad \text{(eq.2)} \tag{7b}$$

If a_{11} and a_{12} are not both zero, then (eq.1) defines a straight line, say $L1$, in a Cartesian x_1, x_2 plane; that is the solution set of (eq.1) is the set of all points on that line. Similarly, if a_{21} and a_{22} are not both zero then the solution set of (eq.2) is the set of all points on a straight line $L2$. There exist exactly three possibilities, and these are illustrated in Fig. 1. First, the lines may intersect at a point, say P, in which case (7) admits the unique solution given by the coordinate pair x_1, x_2 of the point P (Fig. 1a). That is, any solution pair x_1, x_2 of (7) needs to be in the solution set of (eq.1) *and* in the solution set of (eq.2) hence at an intersection of $L1$ and $L2$. This is the generic case, and it occurs (Exercise 2) as long as

$$a_{11}a_{22} - a_{12}a_{21} \neq 0; \tag{8}$$

(8) is the analog of the $a_{11} \neq 0$ condition for the $m = n = 1$ case discussed above.

Second, the lines may be parallel and nonintersecting (Fig. 1b), in which case there is no solution. Then (7) is inconsistent, the solution set is empty.

Third, the lines may coincide (Fig. 1c), in which case the coordinate pair of each point on the line is a solution. Then (7) is consistent and there are an infinite number of solutions.

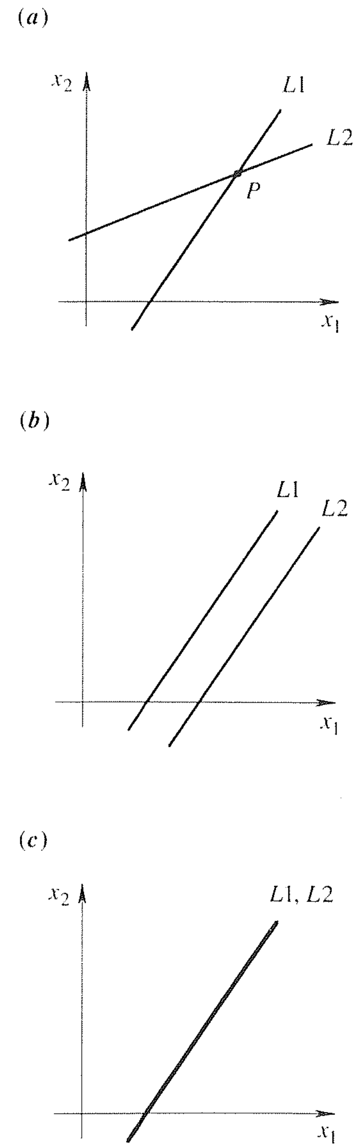


Figure 1. Existence and uniqueness for the system (7).

EXAMPLE 2.

$$\begin{aligned} 2x_1 - x_2 &= 5, \\ x_1 + 3x_2 &= -1, \end{aligned} \qquad \begin{aligned} x_1 + 3x_2 &= 1, \\ x_1 + 3x_2 &= 0, \end{aligned} \qquad \begin{aligned} x_1 + 3x_2 &= 1, \\ 2x_1 + 6x_2 &= 2, \end{aligned}$$

illustrate these three cases, respectively. ▮

Below (7) we said "If a_{11} and a_{12} are not both zero . . . ." What if they are both zero? Then if $c_1 \neq 0$ there is no solution of (7a), and hence there is no solution to the system (7). But if $c_1 = 0$, then (7a) reduces to $0 = 0$ and can be discarded, leaving just (7b). If a_{21} and a_{22} are not both zero, then (7b) gives a line of solutions, but if they are both zero then everything hinges on c_2. If $c_2 \neq 0$ there is no solution and (7) is inconsistent, but if $c_2 = 0$, so both (7a) and (7b) are simply $0 = 0$, then both x_1 and x_2 are arbitrary, and every point in the plane is a solution.

Next, consider the case where $m = n = 3$:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= c_1, && \text{(eq.1)} && (9a) \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= c_2, && \text{(eq.2)} && (9b) \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= c_3. && \text{(eq.3)} && (9c) \end{aligned}$$

Continuing the geometric approach exemplified by Fig. 1, observe that if a_{11}, a_{12}, a_{13} are not all zero then (eq.1) defines a plane, say $P1$, in Cartesian x_1, x_2, x_3 space, and similarly for (eq.2) and (eq.3). In the generic case, $P1$ and $P2$ intersect along a line L, and L pierces $P3$ at a point P. Then the x_1, x_2, x_3 coordinates of P give the unique solution of (9).

In the nongeneric case we can have no solution or an infinity of solutions in the following ways. There will be no solution if L is parallel to $P3$ and hence fails to pierce it, or if any two of the planes are parallel and not coincident. There will be an infinity of solutions if L lies in $P3$ (i.e., a line of solutions), if two planes are coincident and intersect the third (again, a line of solutions), or if all three planes are coincident (this time an entire plane of solutions).

The case where all of the a_{ij} coefficients are zero in one or more of equations (9) is left for the exercises.

An abstract extension of such geometrical reasoning can be continued even if $m = n \geq 4$. For instance, one speaks of $a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = c_1$ as defining a *hyperplane* in an abstract four-dimensional space. In fact, perhaps we should mention that even the familiar x_1, x_2 plane and x_1, x_2, x_3 space discussed here could be abstract as well. For instance, if x_1 and x_2 are unknown currents in two loops of an electrical circuit, then what physical meaning is there to an x_1, x_2 plane? None, but we can introduce it, create it, to assist our reasoning.

Closure. Most of this section is devoted to a geometrical discussion of the system (5) of linear algebraic equations. A great advantage of geometrical reasoning is that it brings our visual system into play. It is estimated that at least a third of the neurons in our brain are devoted to vision, hence our visual sense is extremely sophisticated. No wonder we say "Now I see what you mean; now I get the picture." The more geometry, pictures, visual images to aid our thinking, the better! We have not yet aimed at theorems, and have been content to lay the groundwork for the ideas of existence and uniqueness of solutions. In considering the cases where $m = n = 1$, $m = n = 2$, and $m = n = 3$, we have not meant to imply that we need to have $m = n$; *all* possibilities are considered in the next section. To proceed further, we need to consider the process of *finding* solutions, and that we do, in Section 8.3, by the method of Gauss elimination.

EXERCISES 8.2

1. True or false? If false, give a counterexample.

(a) An algebraic equation is necessarily linear.
(b) An algebraic equation is necessarily nonlinear.
(c) A transcendental equation is necessarily linear.
(d) A transcendental equation is necessarily nonlinear.
(e) A linear equation is necessarily algebraic.
(f) A nonlinear equation is necessarily algebraic.
(g) A linear equation is necessarily transcendental.
(h) A nonlinear equation is necessarily transcendental.

2. Derive the condition (8) as the necessary and sufficient condition for (7) to admit a unique solution.

3. (a) Discuss all possibilities of the existence and uniqueness of solutions of (9) from a geometrical point of view, in the event that $a_{11} = a_{12} = a_{13} = 0$, but a_{21}, a_{22}, a_{23} and a_{31}, a_{32}, a_{33} are not all zero.
(b) Same as (a), but with $a_{21} = a_{22} = a_{23} = 0$ as well.
(c) Same as (a), but with $a_{21} = a_{22} = a_{23} = a_{31} = a_{32} = a_{33} = 0$ as well.

8.3 Solution by Gauss Elimination

8.3.1. Motivation. In this section we continue to consider the system of m linear algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m, \end{aligned} \tag{1}$$

in the n unknowns $x_1, \ldots, x_n$, and develop the solution technique known as **Gauss elimination**. To motivate the ideas, we begin with an example.

EXAMPLE 1. Determine the solution set of the system

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ 3x_1 + x_2 + x_3 &= 9, \\ x_1 - x_2 + 4x_3 &= 8. \end{aligned} \tag{2}$$

Keep the first equation intact, and add -3 times the first equation to the second (as a replacement for the second equation), and add -1 times the first equation to the third (as a replacement for the third equation). These steps yield the new "indented" system

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ -2x_2 + 4x_3 &= 6, \\ -2x_2 + 5x_3 &= 7. \end{aligned} \tag{3}$$

Next, keep the first two of these intact, and add -1 times the second equation to the third, and obtain

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \\ -2x_2 + 4x_3 &= 6, \\ x_3 &= 1. \end{aligned} \tag{4}$$

Finally, multiplying the second of these by $-1/2$ to normalize the leading coefficient (to unity), gives

$$\begin{aligned} x_1 + x_2 - x_3 &= 1, \quad \text{(eq.1)} \\ x_2 - 2x_3 &= -3, \quad \text{(eq.2)} \\ x_3 &= 1. \quad \text{(eq.3)} \end{aligned} \tag{5}$$

It is helpful to think of the original system (2) as a tangle of string that we wish to unravel. The first step is to find a loose end and that is, in effect, what the foregoing process of successive indentations has done for us. Specifically, (eq.3) in (5) is the "loose end," and with that in hand we may unravel (5) just as we would unravel a tangle: putting $x_3 = 1$ into (eq.2) gives $x_2 = -1$, and then putting $x_3 = 1$ and $x_2 = -1$ into (eq.1) gives $x_1 = 3$. Thus, we obtain the unique solution

$$x_3 = 1, \quad x_2 = -1, \quad x_1 = 3. \tag{6}$$

COMMENT 1. From a mathematical point of view, the system (2) was a "tangle" because the equations were **coupled**; that is, each equation contained more than one unknown. Actually, the final system (5) is coupled too, since (eq.1) contains all three unknowns and (eq.2) contains two of them. However, the coupling in (5) is not as debilitating because (5) is in what we call **triangular form**. Thus, we were able to solve (eq.3) for x_3, put that value into (eq.2) and solve for x_2, and then put these values into (eq.1) and solve for x_1, which steps are known as **back substitution**.

COMMENT 2. However, the process begs this question: Is it obvious that the systems (2)–(5) all have the same solution sets so that when we solve (5) we are actually solving (2)? That is, is it not conceivable that in applying the arithmetic steps that carried us from (2) to (5) we might, inadvertently, have altered the solution set? For example, $x - 1 = 4$ has the unique solution $x = 5$, but if we innocently square both sides, the resulting equation $(x-1)^2 = 16$ admits the *two* solutions $x = 5$ and $x = -3$. ▮

The question just raised applies to linear systems in general. It is answered in Theorem 8.3.1 that follows, but first we define two terms: "equivalent systems" and "elementary equation operations."

Two linear systems in n unknowns, x_1 through x_n, are said to be **equivalent** if their solution sets are identical.

The following operations on linear systems are known as **elementary equation operations**:

1. Addition of a multiple of one equation to another
   Symbolically: (eq.j) $\rightarrow$ (eq.j) $+ \alpha$ (eq.k)

2. Multiplication of an equation by a nonzero constant
   Symbolically: (eq.j) $\rightarrow \alpha$ (eq.j)

3. Interchange of two equations
   Symbolically: (eq.j) $\leftrightarrow$ (eq.k)

Then we can state the following result.

THEOREM 8.3.1 *Equivalent Systems*
If one linear system is obtained from another by a finite number of elementary equation operations, then the two systems are equivalent.

Outline of Proof: The truth of this claim for elementary equation operations of types 2 and 3 should be evident, so we confine our remarks to operations of type 1. It suffices to look at the effect of one such operation. Thus, suppose that a given linear system A is altered by replacing its jth equation by its jth plus α times its kth, its other equations being kept intact. Let us call the new system A'. Surely, every solution of A will also be a solution A' since we have merely added equal quantities to equal quantities. That is, *if A' results from A by the application of an elementary equation operation of type 1, then every solution of A is also a solution of A'*. Further, we can convert A' back to A by an elementary equation operation of type 1, namely, by replacing the jth equation of A' by the jth equation of A' plus $-\alpha$ times the kth equation of A'. Consequently, it follows from the italicized result (two sentences back) that every solution of A' is also a solution of A. Then A and A' are equivalent, as claimed. ■

In Example 1, we saw that each step is an elementary equation operation: Three elementary equation operations of type 1 took us from (2) to (4), and one of type 2 took us from (4) to (5); finally, the back substitution amounted to several operations of type 1. Thus, according to Theorem 8.3.1, equivalence was maintained throughout so we can be sure that (6) is the solution set of the original system (2) (as can be verified by direct substitution).

The system in Example 1 admitted a unique solution. To see how the method of successive elimination works out when there is no solution, or a nonunique solution, let us work two more examples.

EXAMPLE 2. *Inconsistent System.* Consider the system

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ x_1 - 2x_2 + x_3 &= 3, \\ 7x_1 \qquad - x_3 &= 2. \end{aligned} \tag{7}$$

Keep the first equation intact, add $-\frac{1}{2}$ times the first equation to the second (eq.2 → eq.2 $-\frac{1}{2}$ eq.1), and add $-\frac{7}{2}$ times the first to the third (eq.3 → eq.3 $-\frac{7}{2}$ eq.1):

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ -\tfrac{21}{2}x_2 + 6x_3 &= -12. \end{aligned} \tag{8}$$

Keep the first two equations intact, and add -3 times the second equation to the third (eq.3

→ eq.3 −3 eq.2):

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ 0 &= -15. \end{aligned} \tag{9}$$

Any solution x_1, x_2, x_3 of (9) must satisfy each of the three equations, but there are no values of x_1, x_2, x_3 that can satisfy $0 = -15$. Thus, (9) is inconsistent (has no solution), and therefore (7) is as well.

COMMENT. The source of the inconsistency is the fact that whereas the left-hand side of the third equation is 2 times the left-hand side of the first equation plus 3 times the left-hand side of the second, the right-hand sides do not bear that relationship: $2(4)+3(3) = 17 \neq 2$. [While that built-in contradiction is not obvious from (7), it eventually comes to light in the third equation in (9).] If we modify the system (7) by changing the final 2 in (7) to 17, then the final -12 in (8) becomes a 3, and the final -15 in (9) becomes a zero:

$$\begin{aligned} 2x_1 + 3x_2 - 2x_3 &= 4, \\ -\tfrac{7}{2}x_2 + 2x_3 &= 1, \\ 0 &= 0 \end{aligned} \tag{10}$$

or, multiplying the first by $\frac{1}{2}$ and the second by $-\frac{2}{7}$,

$$\begin{aligned} x_1 + \tfrac{3}{2}x_2 - x_3 &= 2, \\ x_2 - \tfrac{4}{7}x_3 &= -\tfrac{2}{7}, \end{aligned} \tag{11a,b}$$

where we have discarded the identity $0 = 0$. Thus, by changing the c_j's so as to be "compatible," the system now admits an infinity of solutions rather than none. Specifically, we can let x_3 (or x_2, it doesn't matter which) in (11b) be *any* value, say α, where α is arbitrary. Then (11b) gives $x_2 = -\frac{2}{7} + \frac{4}{7}\alpha$, and putting these into (11a), $x_1 = \frac{17}{7} + \frac{1}{7}\alpha$. Thus, we have the infinity of solutions

$$x_3 = \alpha, \quad x_2 = -\frac{2}{7} + \frac{4}{7}\alpha, \quad x_1 = \frac{17}{7} + \frac{1}{7}\alpha \tag{12}$$

for any α. Evidently, two of the three planes intersect, giving a line that lies in the third plane, and equations (12) are parametric equations of that line! ▮

EXAMPLE 3. *Nonunique Solution.* Consider the system of four equations in six unknowns ($m = 4, n = 6$)

$$\begin{aligned} 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2, \\ x_1 - x_2 + x_3 \qquad\qquad + 2x_6 &= 0, \\ x_1 + x_2 + 2x_3 + 4x_4 + x_5 + 2x_6 &= 3, \\ x_1 - 3x_2 \qquad - 4x_4 - 2x_5 + x_6 &= 0. \end{aligned} \tag{13}$$

Wanting the top equation to begin with x_1 and subsequent equations to indent at the left, let us first move the top equation to the bottom (eq.1 ↔ eq.4):

$$\begin{aligned} x_1 - 3x_2 \quad\quad\; - 4x_4 - 2x_5 + x_6 &= 0, \\ x_1 - x_2 + x_3 \quad\quad\quad\quad\; + 2x_6 &= 0, \\ x_1 + x_2 + 2x_3 + 4x_4 + x_5 + 2x_6 &= 3, \\ 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2. \end{aligned} \tag{14}$$

Add -1 times the first equation to the second (eq.2 → eq.2 -1 eq.1) and third (eq.3 → eq.3 -1 eq.1) equations:

$$\begin{aligned} x_1 - 3x_2 \quad\quad\; - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ 4x_2 + 2x_3 + 8x_4 + 3x_5 + x_6 &= 3, \\ 2x_2 + x_3 + 4x_4 + 3x_5 + x_6 &= 2. \end{aligned} \tag{15}$$

Add -2 times the second to the third (eq.3 → eq.3 -2 eq.2) and -1 times the second to the fourth (eq.4 → eq.4 -1 eq.2):

$$\begin{aligned} x_1 - 3x_2 \quad\quad\; - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ -x_5 - x_6 &= 3, \\ x_5 \quad\quad &= 2. \end{aligned} \tag{16}$$

Add the third to the fourth (eq.4 → eq.4 + eq.3):

$$\begin{aligned} x_1 - 3x_2 \quad\quad\; - 4x_4 - 2x_5 + x_6 &= 0, \\ 2x_2 + x_3 + 4x_4 + 2x_5 + x_6 &= 0, \\ -x_5 - x_6 &= 3, \\ -x_6 &= 5. \end{aligned} \tag{17}$$

Finally, multiply the second, third, and fourth by $\frac{1}{2}$, -1, and -1, respectively, to normalize the leading coefficients (eq.2 → $\frac{1}{2}$ eq.2, eq.3 → -1 eq.3, eq.4 → -1 eq.4):

$$\begin{aligned} x_1 - 3x_2 \quad\quad\; - 4x_4 - 2x_5 + x_6 &= 0, \\ x_2 + \tfrac{1}{2}x_3 + 2x_4 + x_5 + \tfrac{1}{2}x_6 &= 0, \\ x_5 + x_6 &= -3, \\ x_6 &= -5. \end{aligned} \tag{18}$$

The last two equations give $x_6 = -5$ and $x_5 = 2$, and these values can be substituted back into the second equation. In that equation we can let x_4 be arbitrary, say α_1, and we can also let x_3 be arbitrary, say α_2. Then that equation gives x_2 and, again by back substitution, the first equation gives x_1. The result is the infinity of solutions

$$\begin{aligned} x_6 = -5, \quad x_5 = 2, \quad x_4 = \alpha_1, \quad x_3 = \alpha_2, \\ x_2 = \frac{1}{2} - 2\alpha_1 - \frac{1}{2}\alpha_2, \quad x_1 = \frac{21}{2} - 2\alpha_1 - \frac{3}{2}\alpha_2, \end{aligned} \tag{19}$$

where α_1 and α_2 are arbitrary. ■

If a solution set contains p independent arbitrary parameters $(\alpha_1, \ldots, \alpha_p)$, we call it (in this text) a ***p*-parameter family of solutions**. Thus, (12) and (19) are one- and two-parameter families of solutions, respectively. Each choice of values for $\alpha_1, \ldots, \alpha_p$ yields a **particular solution**. In (19), for instance, the choice $\alpha_1 = 1$ and $\alpha_2 = 0$ yields the particular solution $x_1 = \frac{17}{2}$, $x_2 = -\frac{3}{2}$, $x_3 = 0$, $x_4 = 1$, $x_5 = 2$, and $x_6 = -5$.

8.3.2. Gauss elimination. The method of Gauss elimination,* illustrated in Examples 1–3, can be applied to *any* linear system (1), whether or not the system is consistent, and whether or not the solution is unique. Though hard to tell from the foregoing hard calculations, the method is efficient and is commonly available in computer systems.

Observe that the end result of the Gauss elimination process enables us to determine, merely from the pattern of the final equations, whether or not a solution exists and is unique. For instance, we can see from the pattern of (5) that there is a unique solution, from the bottom equation in (9) that there no solution, and from the extra double indentation in (18) that there is a two-parameter family of solutions.

As representative of the case where $m < n$, let $m = 3$ and $n = 5$. There are four possible final patterns, and these are shown schematically in Fig. 1. For instance, the third equation in Fig. 1a could be $x_3 - 6x_4 + 2x_5 = 0$ or $x_3 + 2x_4 + 0x_5 = 4$, and the given third equation in Fig. 1b could be $0 = 6$ or $0 = 0$. It may seem foolish to include the case shown in Fig. 1d because there are no x_j's (all of the a_{ij} coefficients being zero), but it is *possible* so we have included it. From these patterns we draw these conclusions: (a) there exists a two-parameter family of solutions; (b) there is no solution (the system is inconsistent) if the right-hand member of the third equation is nonzero, and a three-parameter family of solutions if the latter is zero; (c) there is no solution if either of the right-hand members of the second and third equations is nonzero, and a four-parameter family of solutions if each of them is zero; (d) there is no solution if any of the right-hand members is nonzero, and a five-parameter family of solutions if each of them is zero.

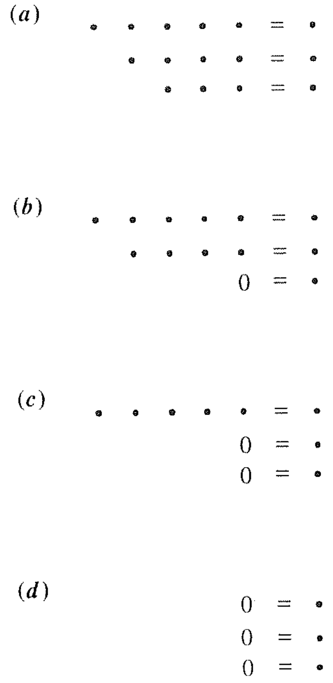


Figure 1. The final pattern; $m = 3, n = 5$.

It may appear that Fig. 1 does not cover all possible cases. For instance, what about the case shown in Fig. 2? That case can be converted to the case shown in Fig. 1a simply by renaming the unknowns: let x_3 become x_2 and let x_5 become x_3. Specifically, let $x_1 \to x_1$, $x_3 \to x_2$, $x_5 \to x_3$, $x_4 \to x_4$, and $x_2 \to x_5$.

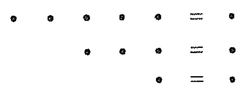


Figure 2. Was this case not covered?

The case where $m \geq n$ can be studied in a similar manner, and we can draw the following general conclusions.

THEOREM 8.3.2 *Existence / Uniqueness for Linear Systems*
If $m < n$, the system (1) can be consistent or inconsistent. If it is consistent it cannot have a unique solution; it will have a p-parameter family of solutions, where $n - m \leq p \leq n$. If $m \geq n$, (1) can be consistent or inconsistent. If it is

*The method is attributed to *Karl Friedrich Gauss* (1777–1855), who is generally regarded as the foremost mathematician of the nineteenth century and often referred to as the "prince of mathematicians."

consistent it can have a unique solution or a p-parameter family of solutions, where $1 \leq p \leq n$.

The next theorem follows immediately from Theorem 8.3.2, but we state it separately for emphasis.

THEOREM 8.3.3 *Existence / Uniqueness for Linear Systems*
Every system (1) necessarily admits no solution, a unique solution, or an infinity of solutions.

Observe that a system (1) is inconsistent only if, in its Gauss-eliminated form, one or more of the equations is of the form zero equal to a nonzero number. But that can never happen if every c_j in (1) is zero, that is, if (1) is **homogeneous**.

THEOREM 8.3.4 *Existence / Uniqueness for Homogeneous Systems*
Every homogeneous linear system of m equations in n unknowns is consistent. Either it admits the unique trivial solution or else it admits an infinity of nontrivial solutions in addition to the trivial solution. If $m < n$, then there is an infinity of nontrivial solutions in addition to the trivial solution.

In summary, not only did the method of Gauss elimination provide us with an efficient and systematic solution procedure, it also led us to important results regarding the existence and uniqueness of solutions.

8.3.3. Matrix notation. In applying Gauss elimination, we quickly discover that writing the variables $x_1, \dots, x_n$ over and over is inefficient, and even tends to upstage the more central role of the a_{ij}'s and c_j's. It is therefore preferable to omit the x_j's altogether and to work directly with the rectangular array

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & c_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & c_m \end{bmatrix}, \tag{20}$$

known as the **augmented matrix** of the system (1), that is, the **coefficient matrix**

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{21}$$

augmented by the column of c_j's. By *matrix* we simply mean a rectangular array of numbers, called **elements**; it is customary to enclose the elements between parentheses to emphasize that the entire matrix is regarded as a single entity. A horizontal line of elements is called a **row**, and a vertical line is called a **column**. Counting rows from the top, and columns from the left,

$$a_{21} \quad a_{22} \quad \cdots \quad a_{2n} \quad c_2 \qquad \text{and} \qquad \begin{matrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{matrix},$$

say, are the second row and $(n+1)$th column, respectively, of the augmented matrix (20).

In terms of the abbreviated matrix notation, the calculation in Example 1 would look like this.

Original system:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 3 & 1 & 1 & 9 \\ 1 & -1 & 4 & 8 \end{bmatrix}.$$

Add -3 times first row to second row, and add -1 times first row to third row:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & -2 & 4 & 6 \\ 0 & -2 & 5 & 7 \end{bmatrix}.$$

Add -1 times second row to third row, and multiply second row by $-\frac{1}{2}$:

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{22}$$

Thus, corresponding to the so-called elementary equation operations on members of a system of linear equations there are **elementary row operations** on the augmented matrix, as follows:

1. Addition of a multiple of one row to another:
   Symbolically: (jth row) → (jth row) + α(kth row)

2. Multiplication of a row by a nonzero constant:
   Symbolically: (jth row) → α(jth row)

3. Interchange of two rows:
   Symbolically: (jth row) ↔ (kth row)

And we say that two matrices are **row equivalent** if one can be obtained from the other by finitely many elementary row operations.

8.3.4. Gauss–Jordan reduction. With the Gauss elimination completed, the remaining steps consist of back substitution. In fact, those steps are elementary row operations as well. The difference is that whereas in the Gauss elimination we proceed from the top down, in the back substitution we proceed from the bottom up.

EXAMPLE 4. To illustrate, let us return to Example 1 and pick up at the end of the Gauss elimination, with (5), and complete the back substitution steps using elementary row operations. In matrix format, we begin with

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 1 & -2 & -3 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{23}$$

Keeping the bottom row intact, add 2 times that row to the second, and add 1 times that row to the first:

$$\begin{bmatrix} 1 & 1 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{24}$$

Now keeping the bottom two rows intact, add -1 times the second row to the first:

$$\begin{bmatrix} 1 & 0 & 0 & 3 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \tag{25}$$

which is the solution: $x_1 = 3$, $x_2 = -1$, $x_3 = 1$ as obtained in Example 1. ▮

The entire process, of Gauss elimination plus back substitution, is known as **Gauss–Jordan reduction**, after Gauss and *Wilhelm Jordan* (1842–1899). The final result is an augmented matrix in **reduced row-echelon form**. That is:

1. In each row not made up entirely of zeros, the first nonzero element is a 1, a so-called **leading 1**.
2. In any two consecutive rows not made up entirely of zeros, the leading 1 in the lower row is to the right of the leading 1 in the upper row.
3. If a column contains a leading 1, every other element in that column is a zero.
4. All rows made up entirely of zeros are grouped together at the bottom of the matrix.

For instance, (25) is in reduced row-echelon form, as is the final matrix in the next example.

EXAMPLE 5. Let us return to Example 3 and finish the Gauss–Jordan reduction, beginning with (18):

$$\begin{bmatrix} 1 & -3 & 0 & -4 & -2 & 1 & 0 \\ 0 & 1 & \frac{1}{2} & 2 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & \frac{3}{2} & 2 & 1 & \frac{5}{2} & 0 \\ 0 & 1 & \frac{1}{2} & 2 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to$$

$$\begin{bmatrix} 1 & 0 & \frac{3}{2} & 2 & 0 & \frac{3}{2} & 3 \\ 0 & 1 & \frac{1}{2} & 2 & 0 & -\frac{1}{2} & 3 \\ 0 & 0 & 0 & 0 & 1 & 1 & -3 \\ 0 & 0 & 0 & 0 & 0 & 1 & -5 \end{bmatrix} \to \begin{bmatrix} \mathbf{1} & 0 & \frac{3}{2} & 2 & 0 & 0 & \frac{21}{2} \\ 0 & \mathbf{1} & \frac{1}{2} & 2 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -5 \end{bmatrix}.$$

The last augmented matrix is in reduced row-echelon form. The four leading 1's are displayed in **bold type**, and we see that, as a result of the back substitution steps, only 0's are to be found above each leading 1. The final augmented matrix once again gives the solution (19). ▮

8.3.5. Pivoting. Recall that the first step in the Gauss elimination of the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m, \end{aligned} \tag{26}$$

is to subtract a_{21}/a_{11} times the first equation from the second, a_{31}/a_{11} times the first equation from the third, and so on, while keeping the first equation intact. The first equation is called the **pivot equation** (or, the first row is the pivot row if one is using the matrix format), and a_{11} is called the **pivot**. That step produces an indented system of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a'_{22}x_2 + \cdots + a'_{2n}x_n &= c'_2, \\ &\vdots \\ a'_{m2}x_2 + \cdots + a'_{mn}x_n &= c'_m. \end{aligned} \tag{27}$$

Next, we keep the first *two* equations intact and use the second equation as the new pivot equation to indent the third through mth equations, and so on.

Naturally, we need each pivot to be nonzero. For instance, we need $a_{11} \neq 0$ for $a_{21}/a_{11}, a_{31}/a_{11}, \ldots$ to be defined. If a pivot is zero, interchange that equation with any one below it, such as the next equation or last equation (as we did in Example 3), until a nonzero pivot is available. Such interchange of equations is called **partial pivoting**. If a pivot is zero we have no choice but to use partial partial pivoting, but in practice even a nonzero pivot should be rejected if it is "very small," since the smaller it is the more susceptible is the calculation to the adverse effect of machine roundoff error (see Exercise 13). To be as safe as possible, one can choose the pivot equation as the one with the largest leading coefficient (relative to the other coefficients in the equation).

Closure. Beginning with a system of coupled linear algebraic equations, one can use a sequence of elementary operations to minimize the coupling between the equations while leaving the solution set intact. Besides putting forward the important method of Gauss elimination, which is used heavily in the following chapters, we used that method to establish several important theoretical results regarding the existence and uniqueness of solutions.

The Gauss elimination and Gauss–Jordan reduction discussions lead naturally to a convenient, and equivalent, formulation in matrix notation. We will return to the concept of matrices in Chapter 10, and develop it in detail.

Computer software. Chapters 8–12 cover the domain known as linear algebra. A great many calculations in linear algebra can be carried out using computer algebra systems. In *Maple*, for instance, a great many commands ("functions") are contained within the **linalg** package. A listing of these commands can be obtained by entering ?linalg. That list includes the **linsolve** command, which can be used to solve a system of m linear algebraic equations in n unknowns. To access linsolve (or any other command within the linalg package), first enter with(linalg). Then, linsolve(A,b) solves (1) for $x_1, \dots, x_n$, where A is the coefficient matrix and b is the column of c_j's. For instance, the system

$$\begin{aligned} x_1 - x_2 + 2x_3 - 3x_4 &= 4, \\ x_1 + 2x_2 - x_3 + 3x_4 &= 1 \end{aligned} \tag{28}$$

admits the two-parameter family of solutions

$$x_4 = \alpha_1, \quad x_3 = \alpha_2, \quad x_2 = -1 - 2\alpha_1 + \alpha_2, \quad x_1 = 3 + \alpha_1 - \alpha_2, \tag{29}$$

where α_1, α_2 are arbitrary. To solve (28) using *Maple*, enter

```
with(linalg):
```

then return and enter

```
linsolve(array([[1, −1, 2, −3], [1, 2, −1, 3]]),  array([4, 1]));
```

and return. The output is

$$[-\_t_1 + \_t_2 + 3, \ \ \_t_1 - 2\_t_2 - 1, \ \ \_t_1, \ \ \_t_2]$$

where the entries are $x_1, \dots, x_4$ and where $\_t_1$ and $\_t_2$ are arbitrary constants. With $\_t_1 = \alpha_2$ and $\_t_2 = \alpha_1$, this result is the same as (29). If you prefer, you could use the sequence

```
with(linalg):
A := array([[1, -1, 2, -3], [1, 2, -1, 3]]):
b := array([4, 1]):
linsolve(A, b);
```

instead. If the system is inconsistent, then either the output will be NULL, or there will be no output.

EXERCISES 8.3

1. Derive the solution set for each of the following systems using Gauss elimination and augmented matrix format. Document each step (e.g., 2nd row→ 2nd row + 5 times 1st row), and classify the result (e.g., unique solution, the system is inconsistent, 3-parameter family of solutions, etc.).

(a)
$$\begin{aligned} 2x - 3y &= 1\\ 5x + y &= 2 \end{aligned}$$

(b)
$$\begin{aligned} 2x + y &= 0\\ 3x - 2y &= 0 \end{aligned}$$

(c) $x + 2y = 4$

(d)
$$\begin{aligned} x - y + z &= 1\\ 2x - y - z &= 8 \end{aligned}$$

(e) $2x_1 - x_2 - x_3 - 5x_4 = 6$

(f)
$$\begin{aligned} 2x_1 - x_2 - x_3 - 3x_4 &= 0\\ x_1 - x_2 + 4x_3 \qquad &= 2 \end{aligned}$$

(g)
$$\begin{aligned} x + 2y + 3z &= 4\\ 5x + 6y + 7z &= 8\\ 9x + 10y + 11z &= 12 \end{aligned}$$

(h)
$$\begin{aligned} x_1 + x_2 - 2x_3 &= 3\\ x_1 - x_2 - 3x_3 &= 1\\ x_1 - 3x_2 - 4x_3 &= -1 \end{aligned}$$

(i)
$$\begin{aligned} 2x_1 - x_2 &= 6\\ 3x_1 + 2x_2 &= 4\\ x_1 + 10x_2 &= -12\\ 6x_1 + 11x_2 &= -2 \end{aligned}$$

(j)
$$\begin{aligned} x_1 - x_2 + 2x_3 + x_4 &= -1\\ 2x_1 + x_2 + x_3 - x_4 &= 4\\ x_1 + 2x_2 - x_3 - 2x_4 &= 5\\ x_1 \qquad + x_3 \qquad &= 1 \end{aligned}$$

(k)
$$\begin{aligned} x_1 \qquad + x_3 \qquad &= 1\\ x_1 + 2x_2 - x_3 - 2x_4 &= 5\\ x_1 - x_2 + 2x_3 + x_4 &= 0\\ 2x_1 + x_2 + x_3 - x_4 &= 4 \end{aligned}$$

(l)
$$\begin{aligned} x_3 + x_4 &= 2\\ 4x_2 - x_3 + x_4 &= 0\\ x_1 - x_2 + 2x_3 + x_4 &= 4 \end{aligned}$$

(m)
$$\begin{aligned} x + 2y + 3z &= 5\\ 2x + 3y + 4z &= 8\\ 3x + 4y + 5z &= c\\ x + y \qquad &= 2 \end{aligned}$$

for $c = 10$, and again, for $c = 11$

(n)
$$\begin{aligned} 2x + y + z &= 10\\ 3x + y - z &= 6\\ x - 2y - 4z &= -10 \end{aligned}$$

(o)
$$\begin{aligned} 2x_1 + x_2 \qquad\qquad &= 1\\ x_1 + 2x_2 + x_3 \qquad &= 1\\ x_2 + 2x_3 + x_4 &= 1\\ x_3 + 2x_4 &= 1 \end{aligned}$$

(p)
$$\begin{aligned} 2x_1 + x_2 \qquad &= 0\\ x_1 + 2x_2 + x_3 &= -1\\ x_2 + 2x_3 &= -4 \end{aligned}$$

(q)
$$\begin{aligned} 2x_1 + x_2 \qquad + x_4 + 2x_5 &= 0\\ x_1 + x_2 - x_3 \qquad\qquad &= 0\\ x_1 + x_2 + x_3 - 3x_4 + 2x_5 &= 0\\ 2x_1 + 2x_2 - x_3 \qquad + x_5 &= 0 \end{aligned}$$

2. (a)–(q) Same as Exercise 1 but using Gauss–Jordan reduction instead of Gauss elimination.

3. (a)–(q) Same as Exercise 1 but using computer software such as the *Maple* linsolve command.

4. Can 20 linear algebraic equations in 14 unknowns have a unique solution? Be inconsistent? Have a two-parameter family of solutions? Have a 14-parameter family of solutions? Have a 16-parameter family of solutions? Explain.

5. Let

$$\begin{aligned} a_1x_1 + a_2x_2 + a_3x_3 &= 0, \\ b_1x_1 + b_2x_2 + b_3x_3 &= 0 \end{aligned}$$

represent any two planes through the origin in a Cartesian x_1, x_2, x_3 space. For the case where the planes intersect along a line, show whether or not that line necessarily passes through the origin.

6. If possible, adapt the methods of this section to solve the following *nonlinear* systems. If it is *not* possible, say so.

(a) $$\begin{aligned} x_1^2 + 2x_2^2 - x_3^2 &= 29 \\ x_1^2 + x_2^2 + x_3^2 &= 19 \\ 3x_1^2 + 4x_2^2 \qquad &= 67 \end{aligned}$$

(b) $$\begin{aligned} x + 3y &= 13 \\ \sin x + 2y &= 5 \end{aligned}$$

(c) $$\begin{aligned} \sin x + \sin y \qquad\qquad &= 1 \\ \sin x - \sin y + 4\cos z &= 1.2 \\ \sin x + \sin y + 2\cos z &= 1.6 \end{aligned}$$

where $-\pi/2 \le x \le \pi/2$, $-\pi/2 \le y \le \pi/2$, and $0 \le z \le 2\pi$.

7. For what values of the parameter λ do the following homogeneous (do you agree that they are homogeneous?) systems admit *non*trivial solutions? Find the nontrivial solutions corresponding to each such λ.

(a) $$\begin{aligned} 2x + y &= \lambda x \\ x + 2y &= \lambda y \end{aligned}$$

(b) $$\begin{aligned} 2x - y &= \lambda x \\ -x + 2y &= \lambda y \end{aligned}$$

(c) $$\begin{aligned} x - 2y &= \lambda x \\ 4x - 8y &= \lambda y \end{aligned}$$

(d) $$\begin{aligned} z &= \lambda x \\ z &= \lambda y \\ x + y + z &= \lambda z \end{aligned}$$

(e) $$\begin{aligned} x + y + z &= \lambda x \\ y + z &= \lambda y \\ 2z &= \lambda z \end{aligned}$$

(f) $$\begin{aligned} 2x + y + z &= \lambda x \\ x + 2y + z &= \lambda y \\ x + y + 2z &= \lambda z \end{aligned}$$

8. Evaluate these excerpts from examination papers.

(a) "Given the system

$$\begin{aligned} x_1 - 2x_2 &= 0, \\ 2x_1 - 4x_2 &= 0, \end{aligned}$$

add -2 times the first equation to the second and add $-\frac{1}{2}$ times the second equation to the first. By these Gauss elimination steps we obtain the equivalent system $0 = 0$ and $0 = 0$, and hence the two-parameter family of solutions $x_1 = \alpha_1$ (arbitrary), $x_2 = \alpha_2$ (arbitrary)."

(b) "Given the system

$$\begin{aligned} x_1 + x_2 - 4x_3 &= 0, \\ 2x_1 - x_2 + x_3 &= 0, \end{aligned}$$

since both left-hand sides equal zero, they must equal each other. Hence we have the equation

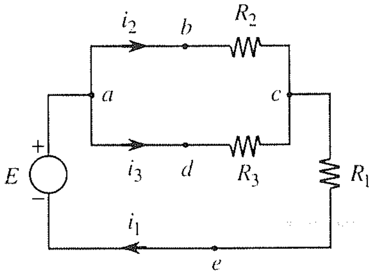
$$x_1 + x_2 - 4x_3 = 2x_1 - x_2 + x_3,$$

which equation is equivalent to the original system."

9. Make up an example of an inconsistent linear algebraic system of equations, with

(a) $m = 2, n = 4$ (b) $m = 1, n = 4$

10. (*Physical example of nonexistence and nonuniqueness; DC circuit*) Kirchoff's current and voltage laws were given in Section 2.3.1. If we apply those laws to the *DC* circuit shown,



we obtain the equations

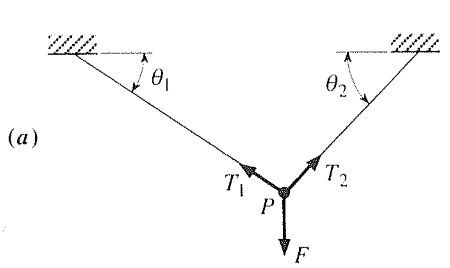
$$\begin{aligned} i_1 - i_2 - i_3 &= 0, && \text{(junction } a\text{)} \\ i_1 - i_2 - i_3 &= 0, && \text{(junction } c\text{)} \\ R_2i_2 - R_3i_3 &= 0, && \text{(loop } abcda\text{)} \\ R_1i_1 + R_2i_2 &= E, && \text{(loop } abcea\text{)} \\ R_1i_1 + R_3i_3 &= E, && \text{(loop } adcea\text{)} \end{aligned} \tag{10.1}$$

where i_1, i_2, i_3 are the three currents (measured as positive in the direction assumed in the figure), R_1, R_2, R_3 are the resistances of the three resistors, and E is the voltage rise (from e to a) induced by the battery or other source. [Evidently, we did not need to apply the current law to both junctions since the resulting equations are identical. Similarly, it may be that not all of the loop equations are needed. But rather than try to decide which of equations (10.1) to keep and which to discard, let us merely keep them all.] We now state the problem: Obtain the solution set of equations (10.1) by Gauss elimination. If there is no solution, or if there is a nonunique solution, explain that result in physical terms. Take

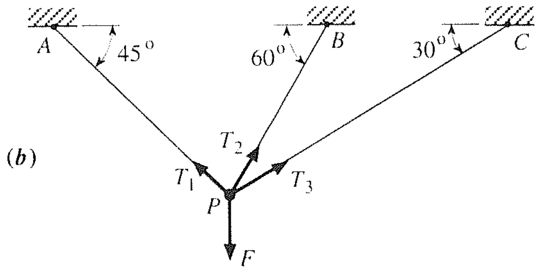
(a) $R_1 = R_2 = R_3 \equiv R \quad (R \neq 0)$
(b) $R_1 = R_2 \equiv R,\ R_2 = 2R \quad (R \neq 0)$
(c) $R_1 \equiv R,\ R_2 = R_3 = 2R \quad (R \neq 0)$

(d) $R_1 \equiv R,\ R_2 = 4R,\ R_3 = 6R \quad (R \neq 0)$
(e) $R_2 \equiv R,\ R_1 = R_3 = 0$
(f) $R_1 \equiv R,\ R_2 = R_3 = 0 \quad (R \neq 0)$
(g) $R_1 = R_2 = R_3 = 0$

11. (*Physical example of nonexistence and nonuniqueness; statically indeterminate structures*) (a) Consider the static equilibrium of the system shown, consisting of two weightless cables connected at P, at which point a vertical load F is applied. Requiring an equilibrium of vertical force components, and horizontal force components too, derive two linear algebraic equations on the unknown tensions T_1 and T_2. Are there any combinations of angles θ_1 and θ_2 (where $0 \leq \theta_1 \leq \frac{\pi}{2}$ and $0 \leq \theta_2 \leq \frac{\pi}{2}$) such that there is either no solution or a nonunique solution? Explain.



(b) This time let there be three cables at angles of 45°, 60°, and 30° as shown. Again, requiring an equilibrium of vertical and horizontal forces at P, derive two linear algebraic equations on the unknown tensions T_1, T_2, T_3. Show that the equations are consistent so there is a nonunique solution. NOTE: We say that such a structure is **statically indeterminate** because the forces in it cannot be determined from the laws of statics alone. What information needs to be added if we are to complete the evaluation of T_1, T_2, T_3? What is needed is information about the relative stiffness of the cables. We pursue this to a conclusion in (c), below.



(c) [*Completion of part (b)*] Before the load F is applied, locate an x, y Cartesian coordinate system at P. Let P be 1 foot below the "ceiling" so the coordinates of A, B, C are $(-1, 1)$, $(1/\sqrt{3}, 1)$, and $(\sqrt{3}, 1)$, respectively. Now apply the load F. The point P will move to a point (x, y), and we assume that the cables are stiff enough so that x and y are very small: $|x| \ll 1$ and $|y| \ll 1$. Let the cables obey Hooke's law: $T_1 = k_1\delta_1$, $T_2 = k_2\delta_2$, and $T_3 = k_3\delta_3$, where δ_j is the increase in length of the jth cable due to the tension T_j. Since P moves to (x, y), it follows that

$$\begin{aligned}\delta_1 &= \sqrt{(x+1)^2 + (y-1)^2} - \sqrt{2}\\ &= \sqrt{2 + 2(x-y) + (x^2+y^2)} - \sqrt{2}\\ &\approx \sqrt{2 + 2(x-y)} - \sqrt{2}\\ &= \sqrt{2}\,[1 + (x-y)]^{1/2} - \sqrt{2}\\ &\approx \sqrt{2}\left[1 + \frac{1}{2}(x-y)\right] - \sqrt{2} = \frac{1}{\sqrt{2}}(x-y).\end{aligned} \tag{11.1}$$

Explain each step in (11.1), and show, similarly, that

$$\delta_2 \approx -\frac{1}{2}x - \frac{\sqrt{3}}{2}y, \tag{11.2}$$

$$\delta_3 \approx -\frac{\sqrt{3}}{2}x - \frac{1}{2}y. \tag{11.3}$$

Thus,

$$\begin{aligned}T_1 &= k_1\delta_1 \approx \frac{k_1}{\sqrt{2}}(x-y),\\ T_2 &= k_2\delta_2 \approx -\frac{k_2}{2}(x + \sqrt{3}y),\\ T_3 &= k_3\delta_3 \approx -\frac{k_3}{2}(\sqrt{3}x + y).\end{aligned} \tag{11.4}$$

Putting (11.4) into the two equilibrium equations obtained in (b) then gives two equations in the unknown displacements x, y. Show that that system can be solved uniquely for x and y, and thus complete the solution for T_1, T_2, T_3.

12. (*Roundoff error difficulty due to small pivots*) To illustrate how small pivots can accentuate the effects of roundoff error, consider the system

$$\begin{aligned}0.005x_1 + 1.47x_2 &= 1.49,\\ 0.975x_1 + 2.32x_2 &= 6.22\end{aligned} \tag{12.1}$$

with exact solution $x_1 = 4$ and $x_2 = 1$. Suppose that our computer carries three significant figures and then rounds off. Using the first equation as our pivot equation, Gauss elimination gives

$$\begin{bmatrix} 0.005 & 1.47 & 1.49 \\ 0.975 & 2.32 & 6.22 \end{bmatrix} \rightarrow \begin{bmatrix} 0.005 & 1.47 & 1.49 \\ 0 & -285 & -284 \end{bmatrix}$$

so $x_2 = 284/285 = 0.996$ and $x_1 = [1.49 - (1.47)(0.996)]/0.005 = (1.49 - 1.46)/0.005 = 6$. Show that if we use partial pivoting and then use the first equation of the system

$$\begin{aligned} 0.975x_1 + 2.32x_2 &= 6.22, \\ 0.005x_1 + 1.47x_2 &= 1.49 \end{aligned} \tag{12.2}$$

as our pivot equation, we obtain the result $x_1 = 4.00$ and $x_2 = 1.00$ (which happens to be exactly correct).

13. (*Ill-conditioned systems*) Practically speaking, our numerical calculations are normally carried out on computers, be they hand-held calculators or large digital computers. Such machines carry only a finite number of significant figures and thus introduce *roundoff error* into most calculations. One might expect (or hope) that such slight deviations will lead to answers that are only slightly in error. For example, the solution of

$$\begin{aligned} x + \quad y &= 2, \\ x - 1.014y &= 0 \end{aligned} \tag{13.1}$$

is $x \approx 1.007$, $y \approx 0.993$, whereas the solution of the rounded-off version

$$\begin{aligned} x + \quad y &= 2, \\ x - 1.01y &= 0 \end{aligned}$$

is very much the same, namely $x \approx 1.005$, $y \approx 0.995$. In sharp contrast, the solutions of

$$\begin{aligned} x + \quad y &= 2, \\ x + 1.014y &= 0 \end{aligned} \tag{13.2}$$

and the rounded-off version

$$\begin{aligned} x + \quad y &= 2, \\ x + 1.01y &= 0, \end{aligned}$$

$x \approx 144.9$, $y \approx -142.9$ and $x = 202$, $y = -200$, respectively; (13.2) is an example of a so-called **ill-conditioned** system (ill-conditioned in the sense that small changes in the coefficients lead to large changes in the solution). Here, we ask the following: Explain why (13.2) is much more sensitive to roundoff than (13.1) by exploring the two cases graphically, that is, in the x, y plane.

Chapter 8 Review

This chapter deals with systems of linear algebraic equations, m equations in n unknowns, insofar as the existence and uniqueness of solutions and solution technique. We find that there are three possibilities: a unique solution, no solution, and an infinity of solutions. If one or more solutions exist then the system is said to be consistent, if there are no solutions then it is inconsistent.

The key, in assessing existence/uniqueness as well as in finding solutions, is provided by elementary operations because they enable us to manipulate the system so as to reduce the coupling to a minimum, while at the same time keeping the solution set intact.

The method of Gauss elimination is introduced, as a systematic solution procedure based upon the three elementary operations, and it is shown that the subsequent back substitution steps amount to elementary operations as well. The entire process, Gauss elimination followed by the back substitution, is known as Gauss–Jordan reduction. Realize that the latter is a solution method, or algorithm, not a formula for the solution. Explicit solution formulas are developed, but not until Chapter 10.

We find that the process of Gauss elimination and Gauss–Jordan reduction are expressed most conveniently in matrix notation, although that notation is not essential to the method. In subsequent chapters the matrix approach is developed more fully.

The most important results of this chapter are contained in Theorems 8.3.1–3. Finally, we also stress the value of geometrical and visual reasoning, and suggest that you keep that idea in mind as we proceed.

Chapter 9

Vector Space

9.1 Introduction

Normally, one meets vectors for the first time within some physical context – in studying mechanics, electric and magnetic fields, and so on. There, the vectors exist within two- or three-dimensional space and correspond to force, velocity, position, magnetic field, and so on. They have both *magnitude* and *direction*; they can be *scaled* by multiplicative factors, *added* according to the parallelogram law; *dot* and *cross product* operations are defined between vectors; the angle between two vectors is defined; vectors can be *expanded* as linear combinations of *base vectors*; and so on.

Alternatively, there exists a highly formalized axiomatic approach to vectors known as *linear vector space* or *abstract vector space*. Although this generalized vector concept is essentially an outgrowth of the more primitive system of "arrow vectors" in 2-space and 3-space, described above, it extends well beyond that system in scope and applicability.

For pedagogical reasons, we break the transition from 2-space and 3-space to abstract vector space into two steps: in Sections 9.4 and 9.5 we introduce a generalization to "n-space," and in Section 9.6 we complete the extension to general vector space, including *function spaces* where the vectors are functions! However, we do not return to function spaces until Chapter 17, in connection with Fourier series and the Sturm–Liouville theory; in Chapters 9–12 our chief interest is in n-space.

9.2 Vectors; Geometrical Representation

Some quantities that we encounter may be completely defined by a single real number, or magnitude; the mass or kinetic energy of a given particle, and the temperature or salinity at some point in the ocean, are examples. Others are not defined solely by a magnitude but rather by a magnitude and a direction, examples being force, velocity, momentum, and acceleration. Such quantities are called

vectors.

The defining features of a vector being magnitude and direction suggests the geometric representation of a vector as a directed line segment, or "arrow," where the length of the arrow is scaled according to the magnitude of the vector. For example, if the wind is blowing at 8 meters/sec from the northeast, that defines a wind-velocity vector $\mathbf{v}$, where we adopt **boldface type** to signify that the quantity is a vector; alternative notations include the use of an overhead arrow as in $\vec{v}$. Choosing, according to convenience, a scale of 5 meters/sec per centimeter, say, the geometric representation of $\mathbf{v}$ is as shown in Fig. 1. Denoting the magnitude, or **norm**, of any vector $\mathbf{v}$ as $\|\mathbf{v}\|$, we have $\|\mathbf{v}\| = 8$ for the $\mathbf{v}$ vector in Fig. 1.

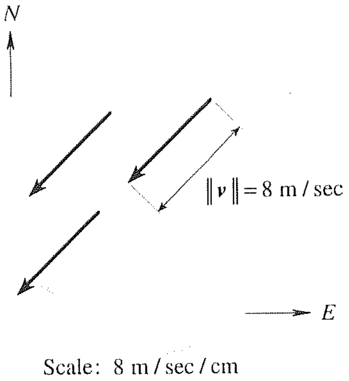


Figure 1. Geometric representation of $\mathbf{v}$.

Observe that the *location* of a vector is not specified, only its magnitude and direction. Thus, the two unlabeled arrows in Fig. 1 are equally valid alternative representations of $\mathbf{v}$. That is not to say that the physical *effect* of the vector will be entirely independent of its position. For example, it should be apparent that the motion of the body $\mathcal{B}$ induced by a force $\mathbf{F}$ (Fig. 2) will certainly depend on the point of application of $\mathbf{F}$[*] as will the stress field induced in $\mathcal{B}$. Nevertheless, the two vectors in Fig. 2 are still regarded as equal, as are the three in Fig. 1.

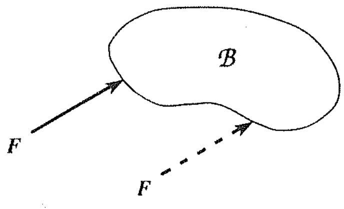


Figure 2. Position of a vector.

Like numbers, vectors do not become useful until we introduce rules for their manipulation, that is, a vector algebra. Having elected the arrow representation of vectors, the vector algebra that we now introduce will, likewise, be geometric.

First, we say that two vectors are **equal** if and only if their lengths are identical and if their directions are identical as well.

Next, we define a process of **addition** between any two vectors $\mathbf{u}$ and $\mathbf{v}$. The first step is to move $\mathbf{v}$ (if necessary), parallel to itself, so that its tail coincides with the head of $\mathbf{u}$. Then the sum, or *resultant*, $\mathbf{u} + \mathbf{v}$ is defined as the arrow from the tail of $\mathbf{u}$ to the head of $\mathbf{v}$, as in Fig. 3a. Reversing the order, $\mathbf{v} + \mathbf{u}$ is as shown in Fig. 3b. Equivalently, we may place $\mathbf{u}$ and $\mathbf{v}$ tail to tail, as in Fig. 3c. Comparing Fig. 3c with Fig. 3a and b, we see that the diagonal of the parallelogram (Fig. 3c), is both $\mathbf{u} + \mathbf{v}$ and $\mathbf{v} + \mathbf{u}$. Thus,

$$\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}, \tag{1}$$

so addition is commutative. One may show (Exercise 3) that it is associative as well,

$$(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}). \tag{2}$$

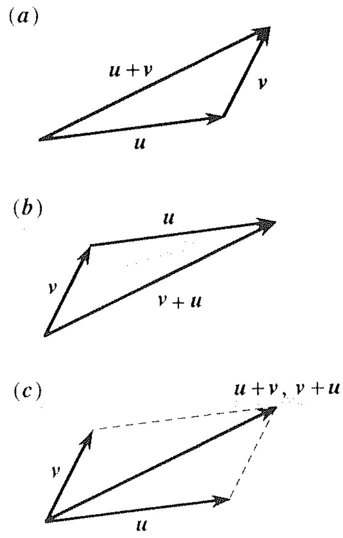


Figure 3. Vector addition.

Next, we define any vector of zero length to be a **zero vector**, denoted as $\mathbf{0}$. Its length being zero, its direction is immaterial; any direction may be assigned if desired. From the definition of addition above, it follows that

$$\mathbf{u} + \mathbf{0} = \mathbf{0} + \mathbf{u} = \mathbf{u} \tag{3}$$

for each vector $\mathbf{u}$.

Corresponding to $\mathbf{u}$ we define a **negative inverse** "$-\mathbf{u}$" such that if $\mathbf{u}$ is any nonzero vector, then $-\mathbf{u}$ is determined uniquely, as shown in Fig. 4a; that is, it is

[*] Students of mechanics know that the point of application of $\mathbf{F}$ affects the *rotational* part of the motion but not the *translational* part.

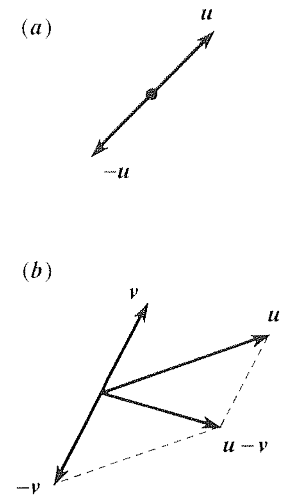


Figure 4. −u and vector subtraction.

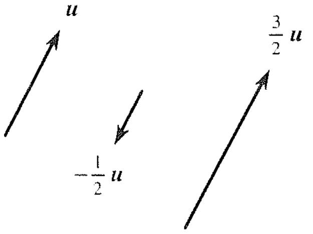


Figure 5. Scalar multiplication.

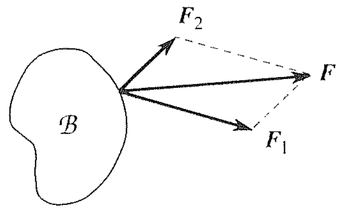


Figure 6. Physical motivation for parallelogram addition.

of the same length as $\mathbf{u}$ but is directed in the opposite direction (again, $\mathbf{u}$ and $-\mathbf{u}$ have the same length, the length of $-\mathbf{u}$ is *not* negative). For the zero vector we have $-\mathbf{0} = \mathbf{0}$. We denote $\mathbf{u} + (-\mathbf{v})$ as $\mathbf{u} - \mathbf{v}$ ("$\mathbf{u}$ minus $\mathbf{v}$") but emphasize that it is really the addition of $\mathbf{u}$ and $-\mathbf{v}$, as in Fig. 4b.

Finally, we introduce another operation, called **scalar multiplication**, between any vector $\mathbf{u}$ and any scalar (i.e., a real number) α: If $\alpha \neq 0$ and $\mathbf{u} \neq \mathbf{0}$, then $\alpha\mathbf{u}$ is a vector whose length is $|\alpha|$ times the length of $\mathbf{u}$ and whose direction is the same as that of $\mathbf{u}$ if $\alpha > 0$, and the opposite if $\alpha < 0$; if $\alpha = 0$ and/or $\mathbf{u} = \mathbf{0}$, then $\alpha\mathbf{u} = \mathbf{0}$. This definition is illustrated in Fig. 5. It follows from this definition that scalar multiplication has the following algebraic properties:

$$\alpha(\beta\mathbf{u}) = (\alpha\beta)\mathbf{u}, \tag{4a}$$

$$(\alpha + \beta)\mathbf{u} = \alpha\mathbf{u} + \beta\mathbf{u}, \tag{4b}$$

$$\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}, \tag{4c}$$

$$1\mathbf{u} = \mathbf{u}, \tag{4d}$$

where α, β are any scalars and $\mathbf{u}, \mathbf{v}$ are any vectors.

Observe that the parallelogram rule of vector addition is a definition so it does not need to be proved. Nevertheless, definitions are not necessarily fruitful so it is worthwhile to reflect for a moment on why the parallelogram rule has proved important and useful. Basically, if we say that "the sum of $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{w}$," and thereby pass from the two vectors $\mathbf{u}, \mathbf{v}$ to the single vector $\mathbf{w}$, it seems fair to expect some sort of equivalence to exist between the action of $\mathbf{w}$ and the joint action of $\mathbf{u}$ and $\mathbf{v}$. For example, if $\mathbf{F}_1$ and $\mathbf{F}_2$ are two forces acting on a body $\mathcal{B}$, as shown in Fig. 6, it is known from fundamental principles of mechanics that their combined effect will be the same as that due to the single force $\mathbf{F}$, so it seems reasonable and natural to say that $\mathbf{F}$ is the sum of $\mathbf{F}_1$ and $\mathbf{F}_2$. This concept goes back at least as far as *Aristotle* (384–322 B.C.). Thus, while the algebra of vectors is developed here as an essentially mathematical matter, it is important to appreciate the role of physics and physical motivation.

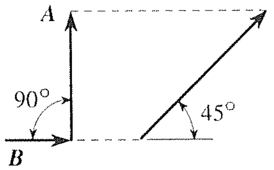
In closing this section, let us remark that our foregoing discussion should not be construed to imply that objects of physical interest are necessarily vectors (as are force and velocity) or scalars [as are temperature, mass, and speed (i.e., the *magnitude of the velocity vector*)]. For example, in the study of mechanics one finds that more than a magnitude and a direction are needed to fully define the state of stress at a point; in fact, a "second-order tensor" is needed – a quantity that is more exotic than a vector in much the same way that a vector is more exotic than a scalar.*

*For an introduction to tensors, we recommend to the interested reader the 68-page book *Tensor Analysis* by H. D. Block (Columbus, OH: Charles E. Merrill, 1962).

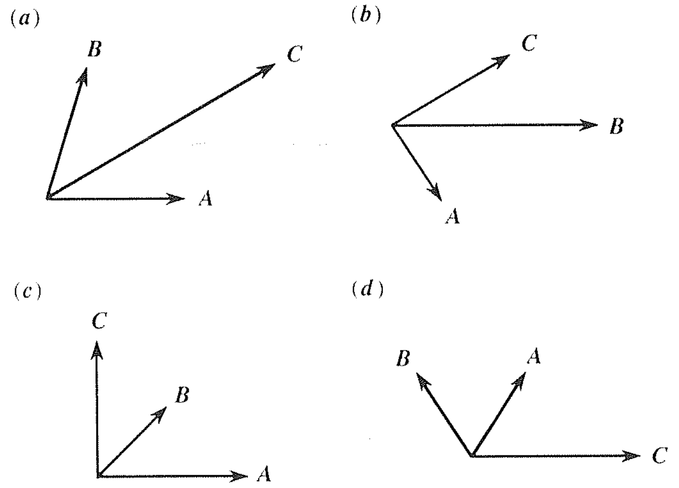
EXERCISES 9.2

1. Trace the vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, shown where $\mathbf{A}$ is twice as long as $\mathbf{B}$. Then determine each of the following by graphical means.

(a) $\mathbf{A}+\mathbf{B}+\mathbf{C}$ (b) $\mathbf{B}-\mathbf{A}$
(c) $\mathbf{A}-\mathbf{C}+3\mathbf{B}$ (d) $2(\mathbf{B}-\mathbf{A})+60$
(e) $\mathbf{A}+(4\mathbf{B}-\mathbf{C})$ (f) $\mathbf{A}+2\mathbf{B}-2\mathbf{C}$



2. In each case, $\mathbf{C}$ can be expressed as a linear combination of $\mathbf{A}$ and $\mathbf{B}$, that is, as $\mathbf{C}=\alpha\mathbf{A}+\beta\mathbf{B}$. Trace the three vectors and by graphical means determine α and β.



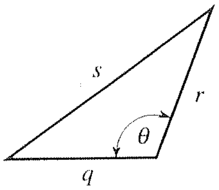
3. Show that the associative property (2) follows from the graphical definition of vector addition.

4. Derive the following from the definitions of vector addition and scalar multiplication:

(a) property (4a) (b) property (4b)
(c) property (4c) (d) property (4d)

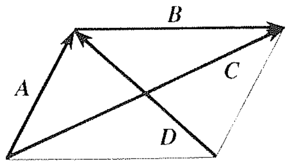
5. (a) If $\|\mathbf{A}\|=1$, $\|\mathbf{B}\|=2$, and $\|\mathbf{C}\|=5$, can $\mathbf{A}+\mathbf{B}+\mathbf{C}=0$?

HINT: Use the *law of cosines* $s^2=q^2+r^2-2qr\cos\theta$ (see the accompanying figure) or the Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides.
(b) Repeat part (a), with "$\|\mathbf{A}\|=1$" changed to $\|\mathbf{A}\|=4$.

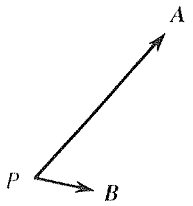


6. Use the definitions and properties given in the reading to show that $\mathbf{A}+\mathbf{B}=\mathbf{C}$ implies that $\mathbf{A}=\mathbf{C}-\mathbf{B}$.

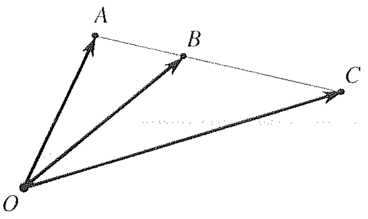
7. (a) Show that if $\mathbf{A}+\mathbf{B}=\mathbf{0}$ and $\mathbf{A}$ and $\mathbf{B}$ are not parallel, then each of $\mathbf{A}$ and $\mathbf{B}$ must be $\mathbf{0}$.
(b) Vectors are often of help in deriving geometrical relationships. For example, to show that the *diagonals of a parallelogram bisect each other* one may proceed as follows. From the accompanying figure $\mathbf{A}+\mathbf{B}=\mathbf{C}$, $\mathbf{A}-\alpha\mathbf{D}=\beta\mathbf{C}$, and $\mathbf{A}=\mathbf{B}+\mathbf{D}$. Eliminating $\mathbf{A}$ and $\mathbf{B}$, we obtain $(2\beta-1)\mathbf{C}=(1-2\alpha)\mathbf{D}$, and since $\mathbf{C}$ and $\mathbf{D}$ are not parallel, it must be true [per part (a)] that $2\beta-1=1-2\alpha=0$ (i.e., $\alpha=\beta=\frac{1}{2}$), which completes the proof. We now state the problem: Use this sort of procedure to show that *a line from one vertex of a parallelogram to the midpoint of a nonadjacent side trisects a diagonal.*



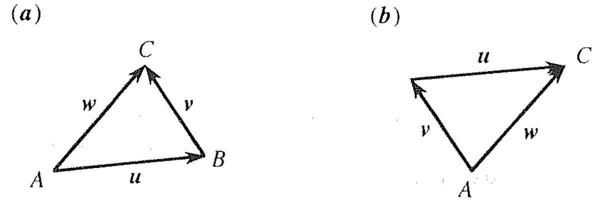
8. If (see the accompanying figure) the vector $\mathbf{A}+\alpha\mathbf{B}$ is placed with its tail at point P, show the line generated by its head as α varies between $-\infty$ and $+\infty$.

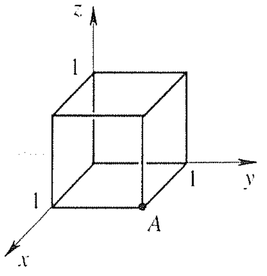


9. If (see the accompanying figure) $\|\mathbf{AB}\|\,/\,\|\mathbf{AC}\|=\alpha$, show that $\mathbf{OB}=\alpha\,\mathbf{OC}+(1-\alpha)\mathbf{OA}$.



10. One may express *linear displacement* as a vector: If a particle moves from point A to point B, the displacement vector is the directed line segment, say $\mathbf{u}$, from A to B. For example, observe that a displacement $\mathbf{u}$ from A to B, followed by a displacement $\mathbf{v}$ from B to C, is equivalent to a single displacement $\mathbf{w}$ from A to C: $\mathbf{u}+\mathbf{v} = \mathbf{w}$ [part (a) in the accompanying figure]. Reversing the order, displacements $\mathbf{v}$ and then $\mathbf{u}$ also carry us from A to C: $\mathbf{v}+\mathbf{u} = \mathbf{w}$ [part (b) in the figure]. Thus, $\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$ so that the commutativity axiom (1) is indeed satisfied. How about *angular* displacements? Suppose that we express the angular displacement of a rigid body about an axis as $\boldsymbol{\theta}$, where the magnitude of $\boldsymbol{\theta}$ is equal to the angle of rotation, and the orientation of $\boldsymbol{\theta}$ is along the axis of rotation, in the direction specified by the "right-hand rule." That is, if we curl the fingers of our right hand about the axis of rotation, in the direction of rotation, then the direction $\boldsymbol{\theta}$ along the axis of rotation is the direction in which our thumb points. The problem is to show that $\boldsymbol{\theta}$, defined in this way, is *not* a proper vector quantity. HINT: Considering the unit cube shown below, say, show (by keeping track of the coordinates of the corner A) that the orientation that results from a rotation of $\pi/2$ about the x axis, followed by a rotation of $\pi/2$ about the y axis, is not the same as that which results when the order of the rotations is reversed. NOTE: If you have encountered angular velocity vectors (usually denoted as ω or $\boldsymbol{\Omega}$), in mechanics, it may seem strange to you that *finite rotations* (assigned a vector direction by the right-hand rule) are not true vectors. The idea is that angular velocity involves *infinitesimal* rotations, and infinitesimal rotations (assigned a vector direction by the right-hand rule) *are* true vectors. This subtle point is discussed in many sources (e.g., Robert R. Long, *Engineering Science Mechanics*, Englewood Cliffs, NJ: Prentice Hall, 1963, pp. 31–36).





9.3 Introduction of Angle and Dot Product

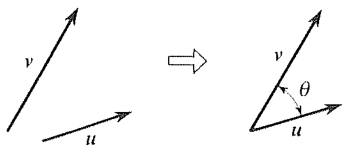


Figure 1. The angle θ between $\mathbf{u}$ and $\mathbf{v}$.

Continuing our discussion, we define here the angle between two vectors and a "dot product" operation between two vectors. The angle θ between two nonzero vectors $\mathbf{u}$ and $\mathbf{v}$ will be understood to mean the ordinary angle between the two vectors when they are arranged tail to tail as in Fig. 1. (We will not attempt to define θ if one or both of the vectors is $\mathbf{0}$.) Of course, this definition of θ is ambiguous in that there are *two* such angles, an interior angle ($\leq \pi$) and an exterior angle ($\geq \pi$); for definiteness, we choose θ to be the interior angle,

$$0 \leq \theta \leq \pi, \tag{1}$$

as in Fig. 1. Unless explicitly stated otherwise, angular measure will be understood to be in *radians*.

Next, we define the so-called **dot product**, $\mathbf{u}\cdot\mathbf{v}$, between two vectors $\mathbf{u}$ and $\mathbf{v}$ as

$$\boxed{\mathbf{u}\cdot\mathbf{v}\equiv\begin{cases}\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta & \text{if } \mathbf{u},\mathbf{v}\neq\mathbf{0},\\ 0 & \text{if } \mathbf{u}=\mathbf{0} \text{ or } \mathbf{v}=\mathbf{0};\end{cases}} \tag{2a,b}$$

$\|\mathbf{u}\|$, $\|\mathbf{v}\|$, and $\cos\theta$ are scalars so $\mathbf{u}\cdot\mathbf{v}$ is a scalar, too.*

By way of geometrical interpretation, observe (Fig. 2a) that $\|\mathbf{u}\|\cos\theta$ is the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$ so that $\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta = (\|\mathbf{v}\|)(\|\mathbf{u}\|\cos\theta)$ is the length of $\mathbf{v}$ times the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$.† Actually, that statement holds if $0\leq\theta\leq\pi/2$; if $\pi/2<\theta\leq\pi$, the cosine is negative, and $\mathbf{u}\cdot\mathbf{v}=\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$ is the negative of the length of $\mathbf{v}$ times the length of the orthogonal projection of $\mathbf{u}$ on the line of action of $\mathbf{v}$.

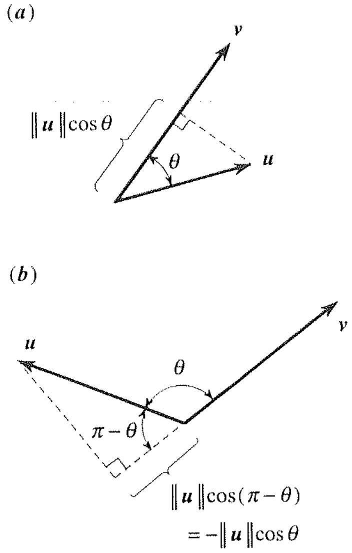


Figure 2. Projection of $\mathbf{u}$ on $\mathbf{v}$.

EXAMPLE 1. *Work Done by a Force.* In mechanics the *work* W done when a body undergoes a linear displacement from an initial point A to a final point B, under the action of a constant force $\mathbf{F}$ (Fig. 3), is defined as the length of the orthogonal projection of $\mathbf{F}$ on the line of displacement, positive if $\mathbf{F}$ is "assisting" the motion (i.e., if $0\leq\theta<\pi/2$, as in Fig. 3a) and negative if $\mathbf{F}$ is "opposing" the motion (i.e., if $\pi/2<\theta<\pi$, as in Fig. 3b), times the displacement. By the displacement we mean the length of the vector $\mathbf{AB}$ with head at B and tail at A. But that product is precisely the dot product of $\mathbf{F}$ with $\mathbf{AB}$,

$$W=\mathbf{F}\cdot\mathbf{AB}. \blacksquare \tag{3}$$

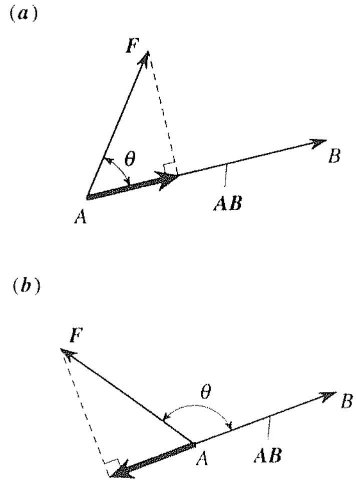


Figure 3. Work done by $\mathbf{F}$.

An important special case of the dot product occurs when $\theta=\pi/2$. Then $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, and

$$\mathbf{u}\cdot\mathbf{v}=\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\frac{\pi}{2}=0. \tag{4}$$

Also of importance is the case where $\mathbf{u}=\mathbf{v}$. Then, according to (2),

$$\mathbf{u}\cdot\mathbf{u}\equiv\begin{cases}\|\mathbf{u}\|\,\|\mathbf{u}\|\cos 0=\|\mathbf{u}\|^2 & \text{if } \mathbf{u}\neq\mathbf{0},\\ 0 & \text{if } \mathbf{u}=\mathbf{0}\end{cases} \tag{5}$$

so that we have

$$\boxed{\|\mathbf{u}\|=\sqrt{\mathbf{u}\cdot\mathbf{u}}} \tag{6}$$

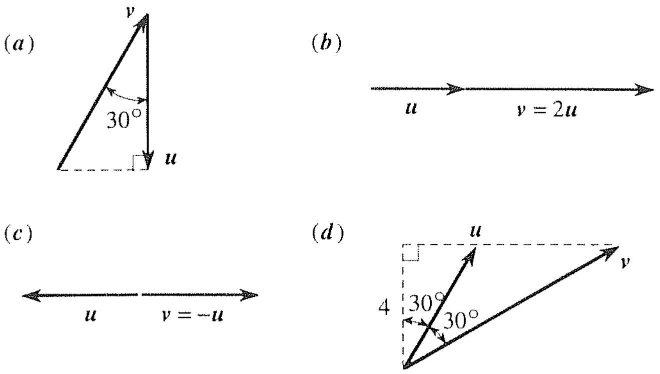
*You may wonder why (2b) is needed since if $\mathbf{u}=\mathbf{0}$, say, then $\|\mathbf{u}\|=0$ and $\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$ is apparently zero anyway. But the point is that if $\mathbf{u}$ and/or $\mathbf{v}$ are $\mathbf{0}$, then θ is undefined; hence $\cos\theta$ (and even zero times $\cos\theta$) is undefined, too.

†Alternatively, we could decompose $\mathbf{u}\cdot\mathbf{v}=\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta=(\|\mathbf{u}\|)(\|\mathbf{v}\|\cos\theta)$; that is, as the length of $\mathbf{u}$ times the length of the orthogonal projection of $\mathbf{v}$ on the line of action of $\mathbf{u}$.

whether $\mathbf{u} \neq \mathbf{0}$ or not. The relationship (6) between the dot product and the norm will be useful in subsequent sections.

EXERCISES 9.3

1. Evaluate $\mathbf{u} \cdot \mathbf{v}$ in each case. In (a) $\|\mathbf{u}\| = 5$, in (b) $\|\mathbf{u}\| = 3$, and in (c) $\|\mathbf{u}\| = 6$.



2. (*Properties of the dot product*) Prove each of the following properties of the dot product, where α, β are any scalars, and $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are any vectors.

(a) $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ (commutativity)

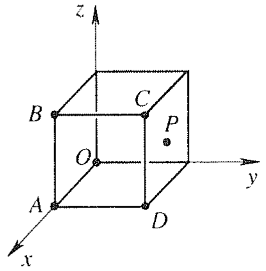
(b) $\mathbf{u} \cdot \mathbf{u} > 0$ for all $\mathbf{u} \neq \mathbf{0}$
$\phantom{(b) \mathbf{u} \cdot \mathbf{u}} = 0$ for all $\mathbf{u} = \mathbf{0}$ (nonnegativeness)

(c) $(\alpha\mathbf{u} + \beta\mathbf{v}) \cdot \mathbf{w} = \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w})$ (linearity)
HINT: In proving part (c), you may wish to show, first, that part (c) is equivalent to the two conditions $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$ and $(\alpha\mathbf{u}) \cdot \mathbf{v} = \alpha(\mathbf{u} \cdot \mathbf{v})$.

3. Using the properties given in Exercise 2, show that

$$(\mathbf{u} + \mathbf{v}) \cdot (\mathbf{w} + \mathbf{x}) = \mathbf{u} \cdot \mathbf{w} + \mathbf{u} \cdot \mathbf{x} + \mathbf{v} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{x}. \qquad (3.1)$$

4. Consider the unit cube shown, where P is the midpoint of the right-hand face. Evaluate each of the following using the definition (2), and (3.1) in Exercise 3. HINT: To evaluate $\mathbf{AC} \cdot \mathbf{OP}$, for instance, write $\mathbf{AC} \cdot \mathbf{OP} = (\mathbf{AD} + \mathbf{DC}) \cdot (\mathbf{OD} + \mathbf{DP})$ and then use (3.1).



(a) $\mathbf{OC} \cdot \mathbf{AB}$ (b) $\mathbf{BA} \cdot \mathbf{OP}$ (c) $\mathbf{AC} \cdot \mathbf{OP}$ (d) $\mathbf{OC} \cdot \mathbf{CP}$
(e) $\mathbf{OC} \cdot \mathbf{OP}$ (f) $\mathbf{BC} \cdot \mathbf{OP}$ (g) $\mathbf{AO} \cdot \mathbf{OP}$ (h) $\mathbf{CP} \cdot \mathbf{DP}$
(i) $\mathbf{BP} \cdot \mathbf{DB}$ (j) $\mathbf{PB} \cdot \mathbf{CO}$ (k) $\mathbf{AP} \cdot \mathbf{PB}$ (l) $\mathbf{AO} \cdot \mathbf{PA}$

5. Referring to the figure in Exercise 4, use the dot product to compute the following angles. (See the hint in Exercise 4.) You may use (2), (6), and (3.1).

(a) APO (b) APB (c) APC (d) APD
(e) ABP (f) ACP (g) BPO (h) BPC
(i) BPD (j) BOP (k) CPO (l) DPO

6. If $\mathbf{u}$ and $\mathbf{v}$ are nonzero, show that $\mathbf{w} = \|\mathbf{v}\|\,\mathbf{u} + \|\mathbf{u}\|\,\mathbf{v}$ bisects the angle between $\mathbf{u}$ and $\mathbf{v}$. (You may use any of the properties given in Exercise 2.)

9.4 n-Space

Here we move away from our dependence on the arrow representation of vectors, in 2-space and 3-space, by introducing an alternative representation in terms of 2-tuples and 3-tuples. This step will lead us to a more general notion of vectors in "n-space."

The idea is simple and is based on the familiar representation of *points* in Cartesian 1-, 2-, and 3-space as 1-, 2-, and 3-tuples of real numbers. For example, the 2-tuple (a_1, a_2) denotes the point P indicated in Fig. 1a, where a_1, a_2 are the x, y coordinates, respectively. But it can also serve to denote the vector $\mathbf{OP}$ in Fig. 1b or, indeed, any equivalent vector $\mathbf{QR}$.

Thus the vector is now represented as the 2-tuple (a_1, a_2) rather than as an arrow, and while pictures may still be drawn, as in Fig. 1b, they are no longer essential and can be discarded if we wish – at least once the algebra of 2-tuples is established (in the next paragraph). The set of all such real 2-tuple vectors will be called **2-space** and will be denoted by the symbol $\mathbb{R}^2$; that is,

$$\mathbb{R}^2 = \{(a_1, a_2) \mid a_1, a_2 \text{ real numbers}\}. \tag{1}$$

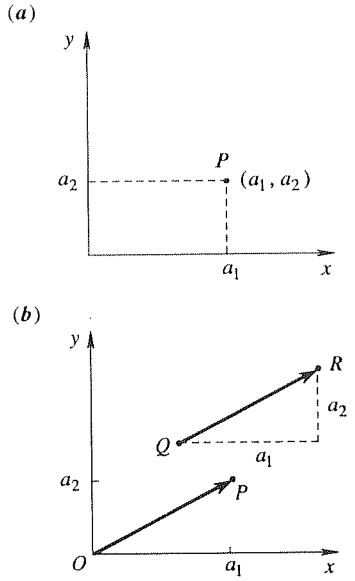


Figure 1. 2-tuple representation.

Vectors $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ in $\mathbb{R}^2$ are defined to be *equal* if $u_1 = v_1$ and $u_2 = v_2$; their *sum* is defined as*

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, u_2 + v_2) \tag{2}$$

as can be seen from Fig. 2; the *scalar multiple* $\alpha\mathbf{u}$ is defined, for any scalar α, as

$$\alpha\mathbf{u} \equiv (\alpha u_1, \alpha u_2); \tag{3}$$

the *zero vector* is

$$\mathbf{0} \equiv (0, 0); \tag{4}$$

and the *negative of* $\mathbf{u}$ is

$$-\mathbf{u} \equiv (-u_1, -u_2). \tag{5}$$

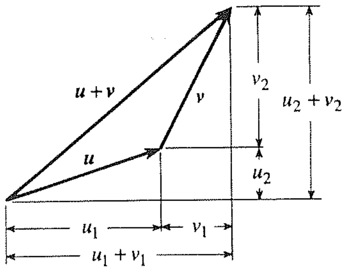


Figure 2. Establishing (2).

Similarly, for $\mathbb{R}^3$:

$$\mathbb{R}^3 = \{(a_1, a_2, a_3) \mid a_1, a_2, a_3 \text{ real numbers}\}. \tag{6}$$

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, u_2 + v_2, u_3 + v_3), \tag{7}$$

and so on.†

It may not be evident that we have gained much since the arrow and n-tuple representations are essentially equivalent. But, in fact, the n-tuple format begins to "open doors." For example, the instantaneous state of the electrical circuit (consisting of a battery and two resistors) shown in Fig. 3 may be defined by the two currents i_1 and i_2 or, equivalently, by the single 2-tuple vector (i_1, i_2). Thus, even though "magnitudes," "directions," and "arrow vectors" may not leap to mind in describing the system shown in Fig. 3, a vector representation *is* quite natural within the n-tuple framework, and that puts us in a position, in dealing with that electrical system, to make use of whatever vector theorems and techniques are available, as developed in subsequent sections and chapters.

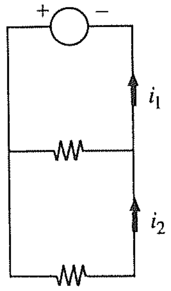


Figure 3. Electrical system.

*We use the ≡ equal sign to mean *equal to by definition*.

†The space $\mathbb{R}^1$ of 1-tuples will not be of interest here.

Indeed, why stop at 3-tuples? One may introduce the set of all ordered real **n-tuple** vectors, even if n is greater than 3. We call this **n-space**, and denote it as $\mathbb{R}^n$, that is,

$$\mathbb{R}^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \text{ real numbers }\}. \tag{8}$$

Consider two vectors, $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$, in $\mathbb{R}^n$. The scalars $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are called the **components** of $\mathbf{u}$ and $\mathbf{v}$. As you may well expect, based on our foregoing discussion of $\mathbb{R}^2$ and $\mathbb{R}^3$, $\mathbf{u}$ and $\mathbf{v}$ are said to be *equal* if $u_1 = v_1, \ldots, u_n = v_n$, and we define

$$\begin{aligned}
\mathbf{u} + \mathbf{v} &\equiv (u_1 + v_1, \ldots, u_n + v_n), && \text{(addition)} && \text{(9a)}\\
\alpha\mathbf{u} &\equiv (\alpha u_1, \ldots, \alpha u_n), && \text{(scalar multiplication)} && \text{(9b)}\\
\mathbf{0} &\equiv (0, \ldots, 0), && \text{(zero vector)} && \text{(9c)}\\
-\mathbf{u} &\equiv (-1)\mathbf{u}, && \text{(negative inverse)} && \text{(9d)}\\
\mathbf{u} - \mathbf{v} &\equiv \mathbf{u} + (-\mathbf{v}). && && \text{(9e)}
\end{aligned}$$

From these definitions we may deduce the following properties:

$$\begin{aligned}
\mathbf{u} + \mathbf{v} &= \mathbf{v} + \mathbf{u}, && \text{(commutativity)} && \text{(10a)}\\
(\mathbf{u} + \mathbf{v}) + \mathbf{w} &= \mathbf{u} + (\mathbf{v} + \mathbf{w}), && \text{(associativity)} && \text{(10b)}\\
\mathbf{u} + \mathbf{0} &= \mathbf{u}, && && \text{(10c)}\\
\mathbf{u} + (-\mathbf{u}) &= \mathbf{0}, && && \text{(10d)}\\
\alpha(\beta\mathbf{u}) &= (\alpha\beta)\mathbf{u}, && \text{(associativity)} && \text{(10e)}\\
(\alpha + \beta)\mathbf{u} &= \alpha\mathbf{u} + \beta\mathbf{u}, && \text{(distributivity)} && \text{(10f)}\\
\alpha(\mathbf{u} + \mathbf{v}) &= \alpha\mathbf{u} + \alpha\mathbf{v}, && \text{(distributivity)} && \text{(10g)}\\
1\mathbf{u} &= \mathbf{u}, && && \text{(10h)}\\
0\mathbf{u} &= \mathbf{0}, && && \text{(10i)}\\
(-1)\mathbf{u} &= -\mathbf{u}, && && \text{(10j)}\\
\alpha\mathbf{0} &= \mathbf{0}. && && \text{(10k)}
\end{aligned}$$

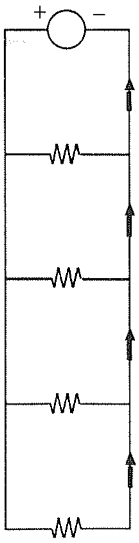


Figure 4. Another electrical circuit.

To illustrate how such n-tuples might arise, observe that the state of the electrical system shown in Fig. 4, may be defined at any instant by the four currents i_1, i_2, i_3, i_4, and that these may be regarded as the components of a single vector $\mathbf{i} = (i_1, i_2, i_3, i_4)$ in $\mathbb{R}^4$.

Of course, the notation of $(u_1, \ldots, u_n)$ as a *point* or *arrow* in an "n-dimensional space" can be realized graphically only if $n \leq 3$; if $n > 3$, the interpretation is valid only in an abstract, or schematic, sense. However, our inability to carry out traditional Cartesian graphical constructions for $n > 3$ will be no hindrance. Indeed, part of the idea here is to move *away* from a dependence on graphical constructions.

Having extended the vector concept to $\mathbb{R}^n$, you may well wonder if further extension is possible. Such extension is not only possible, it constitutes an important step in modern mathematics; more about this in Section 9.6.

EXERCISES 9.4

1. If $\mathbf{t} = (5, 0, 1, 2)$, $\mathbf{u} = (2, -1, 3, 4)$, $\mathbf{v} = (4, -5, 1)$, $\mathbf{w} = (-1, -2, 5, 6)$, evaluate each of the following (as a single vector); if the operation is undefined (i.e., has not been defined here), state that. At each step cite the equation number of the definition or property being used.

(a) $2\mathbf{t} + 7\mathbf{u}$
(b) $3\mathbf{t} - 5\mathbf{u}$
(c) $4[\mathbf{u} + 5(\mathbf{w} - 2\mathbf{u})]$
(d) $4\mathbf{t}\mathbf{u} + \mathbf{w}$
(e) $-\mathbf{w} + \mathbf{t}$
(f) $2\mathbf{t}/\mathbf{u}$
(g) $\mathbf{t} + 2\mathbf{u} + 3\mathbf{w}$
(h) $\mathbf{t} - 2\mathbf{u} - 4\mathbf{v}$
(i) $\mathbf{u}(3\mathbf{t} + \mathbf{w})$
(j) $\mathbf{u}^2 + 2\mathbf{t}$
(k) $2\mathbf{t} + 7\mathbf{u} - 4$
(l) $\mathbf{u} + \mathbf{w}\mathbf{t}$
(m) $\sin \mathbf{u}$
(n) $\mathbf{w} + \mathbf{t} - 2\mathbf{u}$

2. Let $\mathbf{u} = (1, 3, 0, -2)$, $\mathbf{v} = (2, 0, -5, 0)$, and $\mathbf{w} = (4, 3, 2, -1)$.

(a) If $3\mathbf{u} - \mathbf{x} = 4(\mathbf{v} + 2\mathbf{x})$, solve for $\mathbf{x}$ (i.e., find its components).
(b) If $\mathbf{x} + \mathbf{u} + \mathbf{v} + \mathbf{w} = \mathbf{0}$, solve for $\mathbf{x}$.

3. Let $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ be as given in Exercise 2. Citing the definition or property used, at each step, solve each of the following for $\mathbf{x}$. NOTE: Besides the definitions and properties stated in this section, it should be clear that if $\mathbf{x} = \mathbf{y}$, then $\mathbf{x} + \mathbf{z} = \mathbf{y} + \mathbf{z}$ for any $\mathbf{z}$, and $\alpha\mathbf{x} = \alpha\mathbf{y}$ for any α (adding and multiplying equals by equals.)

(a) $3\mathbf{x} + 2(\mathbf{u} - 5\mathbf{v}) = \mathbf{w}$
(b) $3\mathbf{x} = 40 + (1, 0, 0, 0)$
(c) $\mathbf{u} - 4\mathbf{x} = \mathbf{0}$
(d) $\mathbf{u} + \mathbf{v} - 2\mathbf{x} = \mathbf{w}$

4. If $\mathbf{t} = (2, 1, 3)$, $\mathbf{u} = (1, 2, -4)$, $\mathbf{v} = (0, 1, 1)$, $\mathbf{w} = (-2, 1, -1)$, solve each of the following for the scalars $\alpha_1, \alpha_2, \alpha_3$. If no such scalars exist, state that.

(a) $\alpha_1\mathbf{t} + \alpha_2\mathbf{u} + \alpha_3\mathbf{v} = \mathbf{0}$
(b) $\alpha_1\mathbf{t} + \alpha_2\mathbf{v} + \alpha_3\mathbf{w} = \mathbf{0}$
(c) $\alpha_1\mathbf{t} + \alpha_2\mathbf{u} + \alpha_3\mathbf{w} = (1, 3, 2)$
(d) $\alpha_1\mathbf{t} + \alpha_2\mathbf{v} + \alpha_3\mathbf{w} = (2, 0, -1)$
(e) $\alpha_1\mathbf{u} + \alpha_2\mathbf{v} = \mathbf{0}$
(f) $\alpha_1\mathbf{u} + \alpha_2\mathbf{v} = \alpha_3\mathbf{w} - (2, 0, 0)$

5. (a) If $\mathbf{u}$ and $\mathbf{v}$ are given 4-tuples and $\mathbf{0} = (0, 0, 0, 0)$, does the vector equation $\alpha_1\mathbf{u} + \alpha_2\mathbf{v} = \mathbf{0}$ necessarily have nontrivial solutions for the scalars α_1 and α_2? Explain. (If the answer is "no," a counterexample will suffice.)
(b) Repeat part (a), but where $\mathbf{u}$, $\mathbf{v}$ are 3-tuples and $\mathbf{0} = (0, 0, 0)$.
(c) Repeat part (a), but where $\mathbf{u}$, $\mathbf{v}$ are 2-tuples and $\mathbf{0} = (0, 0)$.

9.5 Dot Product, Norm, and Angle for n-Space

9.5.1. Dot product, norm, and angle. We wish to define the norm of an n-tuple vector, and the dot product and angle between two n-tuple vectors, just as we did for "arrow vectors." These definitions should be expressed in terms of the *components* of the n-tuples since the graphical and geometrical arguments used for arrow vectors will not be possible here for $n > 3$. Thus, if $\mathbf{u} = (u_1, \dots, u_n)$, we wish to define the *norm* or "length" of $\mathbf{u}$, denoted as $\|\mathbf{u}\|$, in terms of the components $u_1, \dots, u_n$ of $\mathbf{u}$; and given another vector $\mathbf{v} = (v_1, \dots, v_n)$, we wish to define the *angle* θ between $\mathbf{u}$ and $\mathbf{v}$, and the *dot product* $\mathbf{u} \cdot \mathbf{v}$, in terms of $u_1, \dots, u_n$ and $v_1, \dots, v_n$.

Furthermore, we would like these definitions to reduce to the definitions given in Sections 9.2 and 9.3 in the event that $n = 2$ or 3.

Let us begin with the dot product. Our plan is to return to the arrow vector formula

$$\mathbf{u} \cdot \mathbf{v} = \|\mathbf{u}\| \, \|\mathbf{v}\| \cos\theta, \tag{1}$$

to re-express it in terms of vector components for $\mathbb{R}^2$ and $\mathbb{R}^3$, and then to generalize those forms to $\mathbb{R}^n$.

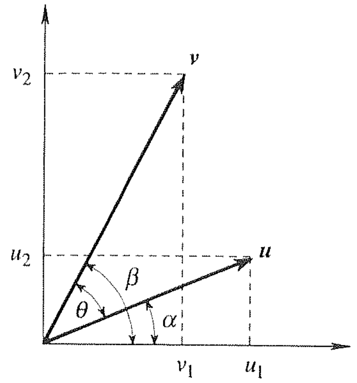


Figure 1. $\mathbf{u}\cdot\mathbf{v}$ in terms of components.

If $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbb{R}^2$ as shown in Fig. 1, formula (1) may be expressed in terms of the components of $\mathbf{u}$ and $\mathbf{v}$ as follows:

$$\begin{aligned}\mathbf{u}\cdot\mathbf{v} &= \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta \\ &= \|\mathbf{u}\|\,\|\mathbf{v}\|\cos(\beta-\alpha) \\ &= \|\mathbf{u}\|\,\|\mathbf{v}\|\,(\cos\beta\cos\alpha+\sin\beta\sin\alpha) \\ &= (\|\mathbf{u}\|\cos\alpha)\,(\|\mathbf{v}\|\,(\cos\beta)+(\|\mathbf{u}\|\sin\alpha)\,(\|\mathbf{v}\|\,(\sin\beta) \\ &= u_1v_1+u_2v_2. \end{aligned} \tag{2}$$

We state, without derivation, that the analogous result for $\mathbb{R}^3$ is

$$\mathbf{u}\cdot\mathbf{v} = u_1v_1+u_2v_2+u_3v_3. \tag{3}$$

Generalizing (2) and (3) to $\mathbb{R}^n$, it is eminently reasonable to define the (scalar-valued) dot product of two n-tuple vectors $\mathbf{u} = (u_1,\dots,u_n)$ and $\mathbf{v} = (v_1,\dots,v_n)$ as

$$\boxed{\mathbf{u}\cdot\mathbf{v} \equiv u_1v_1+u_2v_2+\cdots+u_nv_n = \sum_{j=1}^{n} u_jv_j.} \tag{4}$$

Observe that we have not proved (4); it is a definition.

Defining the dot product is the key, for now $\|\mathbf{u}\|$ and θ follow readily. Specifically, we define

$$\boxed{\|\mathbf{u}\| \equiv \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j^2}} \tag{5}$$

in accordance with equation (6) in Section 9.3, and

$$\boxed{\theta \equiv \cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right),} \tag{6}$$

from (1), where the inverse cosine is understood to be in the interval $[0,\pi]$.* Notice the generalized Pythagorean-theorem nature of (5).

Other dot products and norms are sometimes defined for n-space, but we choose to use (4) and (5), which are known as the **Euclidean dot product** and **Euclidian norm**, respectively. To signify that the Euclidean dot product and norm have been adopted, we henceforth refer to the space as **Euclidean n-space**, rather

*By the "interval $[a, b]$ on a real x axis," we mean the points $a \le x \le b$. Such an interval is said to be **closed** since it includes the two endpoints. To denote the **open** interval $a < x < b$, we write (a, b). Similarly, $[a, b)$ means $a \le x < b$, and $(a, b]$ means $a < x \le b$. Implicit in the closed-interval notation $[a, b]$ is the finiteness of a and b.

than just n-space. We will still denote it by the symbol $\mathbb{R}^n$ (although some authors prefer the notation $\mathbb{E}^n$).

EXAMPLE 1. Let $\mathbf{u} = (1, 0)$ and $\mathbf{v} = (2, -2)$. Then

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= (1)(2) + (0)(-2) = 2, \\ \|\mathbf{u}\| &= \sqrt{(1)^2 + (0)^2} = 1, \\ \|\mathbf{v}\| &= \sqrt{(2)^2 + (-2)^2} = 2\sqrt{2}, \\ \theta &= \cos^{-1}\left(\frac{2}{2\sqrt{2}}\right) = \frac{\pi}{4} \quad (\text{or } 45^\circ), \end{aligned}$$

as is readily verified if we sketch $\mathbf{u}$ and $\mathbf{v}$ as arrow vectors in a Cartesian plane. ▮

EXAMPLE 2. Let $\mathbf{u} = (2, -2, 4, -1)$ and $\mathbf{v} = (5, 9, -1, 0)$. Then,

$$\mathbf{u} \cdot \mathbf{v} = (2)(5) + (-2)(9) + (4)(-1) + (-1)(0) = -12, \tag{7}$$

$$\|\mathbf{u}\| = \sqrt{(2)^2 + (-2)^2 + (4)^2 + (-1)^2} = 5, \tag{8}$$

$$\|\mathbf{v}\| = \sqrt{(5)^2 + (9)^2 + (-1)^2 + (0)^2} = \sqrt{107}, \tag{9}$$

$$\theta = \cos^{-1}\left(\frac{-12}{5\sqrt{107}}\right) \approx \cos^{-1}(-0.232) \approx 1.805 \quad (\text{or } 103.4^\circ). \tag{10}$$

In this case, n $(= 4)$ is greater than 3 so (7) through (10) are not to be understood in any physical or graphical sense, but merely in terms of the definitions (4) to (6).

COMMENT. The dot product of $\mathbf{u} = (2, -2, 4)$ and $\mathbf{v} = (5, 9, -1, 0)$, on the other hand, is *not defined* since here $\mathbf{u}$ and $\mathbf{v}$ are members of different spaces, $\mathbb{R}^3$ and $\mathbb{R}^4$, respectively. It is *not* legitimate to augment $\mathbf{u}$ to the form $(2, -2, 4, 0)$ on the grounds that "surely adding a zero can't hurt." ▮

There is one catch that you may have noticed: (6) serves to define a (real) θ only if the argument of the inverse cosine is less than or equal to unity in magnitude. That this is indeed true is not so obvious. Nevertheless, that

$$-1 \leq \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \leq 1 \qquad \text{or} \qquad |\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\| \tag{11}$$

does necessarily hold will be proved in a moment. Whereas double braces denote vector norm, the single braces in (11) denote the absolute value of the scalar $\mathbf{u} \cdot \mathbf{v}$.

9.5.2. Properties of the dot product. The dot product defined by (4) possesses the following important properties:

$$\begin{aligned} &\textit{Commutative}: & \mathbf{u} \cdot \mathbf{v} &= \mathbf{v} \cdot \mathbf{u}, & &\text{(12a)} \\ &\textit{Nonnegative}: & \mathbf{u} \cdot \mathbf{u} &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0} & & \\ & & &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, & &\text{(12b)} \\ &\textit{Linear}: & (\alpha\mathbf{u} + \beta\mathbf{v}) \cdot \mathbf{w} &= \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w}), & &\text{(12c)} \end{aligned}$$

for any scalars α, β and any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$. The linearity condition (12c) is equivalent to the two conditions $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = (\mathbf{u}\cdot\mathbf{w})+(\mathbf{v}\cdot\mathbf{w})$ and $(\alpha\mathbf{u})\cdot\mathbf{v} = \alpha(\mathbf{u}\cdot\mathbf{v})$. Verification of these claims is left for the exercises.

EXAMPLE 3. Expand the dot product $(6\mathbf{t}-2\mathbf{u})\cdot(\mathbf{v}+4\mathbf{w})$. Using (12), we obtain

$$\begin{aligned}(6\mathbf{t}-2\mathbf{u})\cdot(\mathbf{v}+4\mathbf{w}) &= 6[\mathbf{t}\cdot(\mathbf{v}+4\mathbf{w})] - 2[\mathbf{u}\cdot(\mathbf{v}+4\mathbf{w})] && \text{by (12c)}\\ &= 6[(\mathbf{v}+4\mathbf{w})\cdot\mathbf{t}] - 2[(\mathbf{v}+4\mathbf{w})\cdot\mathbf{u}] && \text{by (12a)}\\ &= 6(\mathbf{v}\cdot\mathbf{t}) + 24(\mathbf{w}\cdot\mathbf{t}) - 2(\mathbf{v}\cdot\mathbf{u}) - 8(\mathbf{w}\cdot\mathbf{u}) && \text{by (12c)}\end{aligned}$$

in much the same way that we obtain $(a-b)(c+d) = ac+ad-bc-bd$ in scalar arithmetic. ▮

As a consequence of (12) we are in a position to prove the promised inequality (11), namely, the **Schwarz inequality**[*]

$$\boxed{|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|\,.} \tag{13}$$

To derive this result, we start with the inequality

$$(\mathbf{u}+\alpha\mathbf{v})\cdot(\mathbf{u}+\alpha\mathbf{v}) \ge 0, \tag{14}$$

which is guaranteed by (12b), for any scalar α and any vectors $\mathbf{u}$ and $\mathbf{v}$. Expanding the left-hand side and noting that $\mathbf{u}\cdot\mathbf{u} = \|\mathbf{u}\|^2$ and $\mathbf{v}\cdot\mathbf{v} = \|\mathbf{v}\|^2$, (14) becomes[†]

$$\|\mathbf{u}\|^2 + 2\alpha\mathbf{u}\cdot\mathbf{v} + \alpha^2\|\mathbf{v}\|^2 \ge 0. \tag{15}$$

Regarding $\mathbf{u}$ and $\mathbf{v}$ as fixed and α as variable, the left-hand side is then a quadratic function of α. If we choose α so as to minimize the left-hand side, then (15) will be as close to an equality as possible and hence as informative as possible. Thus, setting $d(\text{left-hand side})/d\alpha = 0$, we obtain

$$2\mathbf{u}\cdot\mathbf{v} + 2\alpha\|\mathbf{v}\|^2 = 0 \qquad \text{or} \qquad \alpha = -\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{v}\|^2}.$$

Putting this optimal value of α back into (15) gives us

$$\begin{aligned}\|\mathbf{u}\|^2 - 2\frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2} + \frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2} &\ge 0,\\ \|\mathbf{u}\|^2\|\mathbf{v}\|^2 - 2(\mathbf{u}\cdot\mathbf{v})^2 + (\mathbf{u}\cdot\mathbf{v})^2 &\ge 0,\\ \|\mathbf{u}\|^2\|\mathbf{v}\|^2 &\ge (\mathbf{u}\cdot\mathbf{v})^2,\end{aligned}$$

[*] After *Hermann Amandus Schwarz* (1843–1921). The names *Cauchy* and *Bunyakovsky* are also associated with this well-known inequality.

[†] Does a term such as $\alpha\mathbf{u}\cdot\mathbf{v}$ in (15) mean $(\alpha\mathbf{u})\cdot\mathbf{v}$, or $\alpha(\mathbf{u}\cdot\mathbf{v})$? It does not matter; by virtue of (12c) (with $\beta = 0$ and $\mathbf{w}$ changed to $\mathbf{v}$), $(\alpha\mathbf{u})\cdot\mathbf{v} = \alpha(\mathbf{u}\cdot\mathbf{v})$, so the parentheses are not needed.

and taking square roots of both sides yields the Schwarz inequality (13).*

Thus, it was not merely a matter of luck that the arguments of the inverse cosines were smaller than unity in magnitude in Examples 1 and 2, it was guaranteed in advance by the Schwarz inequality (13).

9.5.3. Properties of the norm. Since the norm is related to the dot product according to

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}, \tag{16}$$

the properties (12) of the dot product should imply certain corresponding properties of the norm. These properties are as follows:

$$\textit{Scaling:} \qquad \|\alpha\mathbf{u}\| = |\alpha|\,\|\mathbf{u}\|\,, \tag{17a}$$

$$\textit{Nonnegative:} \qquad \begin{aligned} \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u}\neq\mathbf{0} \\ &= 0 \quad \text{for } \mathbf{u}=\mathbf{0}, \end{aligned} \tag{17b}$$

$$\textit{Triangle Inequality:} \qquad \|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\| + \|\mathbf{v}\|\,. \tag{17c}$$

Equation (17a) simply says that $\alpha\mathbf{u}$ is $|\alpha|$ times as long as $\mathbf{u}$, and for arrow representations of 2-tuples or 3-tuples the triangle inequality (17c) amounts to the Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides (Fig. 2). Less obvious, however, is the fact that (17c) holds for n-tuples for n's > 3.

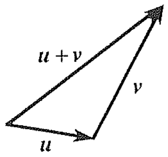


Figure 2. Triangle inequality.

Let us prove only (17a) and (17c) since (17b) follows readily from (16) and (12b). First, (17a):

$$\begin{aligned} \|\alpha\mathbf{u}\| &= \sqrt{(\alpha\mathbf{u})\cdot(\alpha\mathbf{u})} && \text{by (16)} \\ &= \sqrt{\alpha\mathbf{u}\cdot(\alpha\mathbf{u})} && \text{by (12c) with } \beta=0 \text{ and } \mathbf{w}=\alpha\mathbf{u} \\ &= \sqrt{\alpha(\alpha\mathbf{u})\cdot\mathbf{u}} && \text{by (12a)} \\ &= \sqrt{\alpha^2\mathbf{u}\cdot\mathbf{u}} && \text{by (12c) with } \beta=0 \text{ and } \mathbf{w}=\mathbf{u} \\ &= |\alpha|\sqrt{\mathbf{u}\cdot\mathbf{u}} = |\alpha|\,\|\mathbf{u}\|\,. \end{aligned}$$

Turning to (17c), we find that

$$\begin{aligned} \|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) && \text{by (16)} \\ &= \mathbf{u}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{u}+\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{v} \\ &= \|\mathbf{u}\|^2 + 2\mathbf{u}\cdot\mathbf{v} + \|\mathbf{v}\|^2 \\ &\le \|\mathbf{u}\|^2 + 2\,|\mathbf{u}\cdot\mathbf{v}| + \|\mathbf{v}\|^2 \\ &\le \|\mathbf{u}\|^2 + 2\,\|\mathbf{u}\|\,\|\mathbf{v}\| + \|\mathbf{v}\|^2 && \text{by (13)} \\ &= (\|\mathbf{u}\| + \|\mathbf{v}\|)^2 \end{aligned}$$

*That the choice $\alpha = -\mathbf{u}\cdot\mathbf{v}/\|\mathbf{v}\|^2$ *min*imizes the left-hand side of (15) follows from the fact that $d^2(\text{left-hand side})/d\alpha^2 = 2\|\mathbf{v}\|^2 > 0$.

so that

$$\|\mathbf{u}+\mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|,$$

as claimed. A key step was the use of the Schwarz inequality (13), but we also used the simple inequality $\mathbf{u}\cdot\mathbf{v} \leq |\mathbf{u}\cdot\mathbf{v}|$, which holds since $\mathbf{u}\cdot\mathbf{v}$ is a (positive, zero, or negative) real number; that is, if $\mathbf{u}\cdot\mathbf{v}$ is negative, then the $<$ holds, and if $\mathbf{u}\cdot\mathbf{v}$ is zero or positive, then the $=$ holds.

EXAMPLE 4. Let us verify the triangle inequality for a specific example, say the vectors $\mathbf{u} = (2,1,3,-1)$ and $\mathbf{v} = (0,4,2,1)$. Then $\mathbf{u}+\mathbf{v} = (2,5,5,0)$ so (17c) becomes

$$\sqrt{54} \leq \sqrt{15} + \sqrt{21}$$

or $7.348 \leq 3.873 + 4.583$, which is indeed true. ▮

9.5.4. Orthogonality. If $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors such that $\mathbf{u}\cdot\mathbf{v} = 0$, then

$$\theta = \cos^{-1}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right) = \cos^{-1}\left(\frac{0}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right) = \cos^{-1}(0) = \frac{\pi}{2}, \tag{18}$$

and we say that $\mathbf{u}$ and $\mathbf{v}$ are *perpendicular*. [Here we have used the nonzeroness of $\mathbf{u}$ and $\mathbf{v}$ in the third equality in (18); if $\mathbf{u}$ and/or $\mathbf{v}$ *were* $\mathbf{0}$, we would have had $\cos^{-1}(0/\|\mathbf{u}\|\,\|\mathbf{v}\|) = \cos^{-1}(0/0)$, which is not defined.]

But to equate the condition $\mathbf{u}\cdot\mathbf{v} = 0$ to perpendicularity ($\theta = \pi/2$) would not be correct since $\mathbf{u}\cdot\mathbf{v}$ will also be zero in the event that $\mathbf{u}$ and/or $\mathbf{v}$ are $\mathbf{0}$, in which case θ is not defined. Let us therefore make a distinction between perpendicularity and "orthogonality." We will say that $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if

$$\mathbf{u}\cdot\mathbf{v} = 0. \tag{19}$$

Only if $\mathbf{u}$ and $\mathbf{v}$ are both nonzero does their orthogonality imply their perpendicularity (i.e., $\theta = \pi/2$). With this definition, we see that the zero vector $\mathbf{0}$ is orthogonal to *every* vector including itself (Exercise 14).

Finally, we say that a *set* of vectors, say $\{\mathbf{u}_1,\ldots,\mathbf{u}_k\}$, is an **orthogonal set** if every vector in the set is orthogonal to every other one:

$$\mathbf{u}_i\cdot\mathbf{u}_j = 0 \quad \text{if } i \neq j. \tag{20}$$

EXAMPLE 5. $\mathbf{u}_1 = (2,3,-1,0)$, $\mathbf{u}_2 = (1,2,8,3)$, $\mathbf{u}_3 = (9,-6,0,1)$ is an orthogonal set because $\mathbf{u}_1\cdot\mathbf{u}_2 = \mathbf{u}_1\cdot\mathbf{u}_3 = \mathbf{u}_2\cdot\mathbf{u}_3 = 0$. ▮

EXAMPLE 6. $\mathbf{u}_1 = (1,3)$, $\mathbf{u}_2 = (0,0)$ is an orthogonal set because $\mathbf{u}_1\cdot\mathbf{u}_2 = 0$. ▮

9.5.5. Normalization. Any nonzero vector $\mathbf{u}$ can be scaled to have unit length by multiplying it by $1/\,\|\mathbf{u}\|$ so we say that the vector

$$\hat{\mathbf{u}} = \frac{1}{\|\mathbf{u}\|}\mathbf{u} \tag{21}$$

has been "normalized." That $\hat{\mathbf{u}}$ has unit length is readily verified:

$$\begin{aligned} \|\hat{\mathbf{u}}\| = \left\|\frac{1}{\|\mathbf{u}\|}\mathbf{u}\right\| &= \left|\frac{1}{\|\mathbf{u}\|}\right| \|\mathbf{u}\| && \text{by (17a)} \\ &= \frac{1}{\|\mathbf{u}\|}\|\mathbf{u}\| && \text{by (17b)} \\ &= 1. \end{aligned}$$

A vector of unit length is called a **unit vector**. We will often use the caret notation $\hat{\mathbf{u}}$ for unit vectors.

EXAMPLE 7. Normalize $\mathbf{u} = (1, -1, 0, 2)$. Since $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{6}$, we have

$$\hat{\mathbf{u}} = \frac{1}{\|\mathbf{u}\|}\mathbf{u} = \frac{1}{\sqrt{6}}(1, -1, 0, 2) = \left(\frac{1}{\sqrt{6}}, -\frac{1}{\sqrt{6}}, 0, \frac{2}{\sqrt{6}}\right). \quad ▮$$

A set of vectors is said to be **orthonormal** if it is orthogonal and if each vector is normalized (i.e., is a unit vector). We will use that term so frequently that it will be useful to abbreviate it as ON, but be aware that that abbreviation is not standard. Thus, $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is ON if and only if $\mathbf{u}_i \cdot \mathbf{u}_j = 0$ whenever $i \neq j$ (for orthogonality), and $\mathbf{u}_j \cdot \mathbf{u}_j = 1$ for each j (so $\|\mathbf{u}_j\| = 1$, so the set is normalized).

The symbol

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{22}$$

will be useful, and is known as the **Kronecker delta**, after *Leopold Kronecker* (1823–1891). Thus, $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is ON if and only if

$$\boxed{\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij}} \tag{23}$$

for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k$.

EXAMPLE 8. Let

$$\mathbf{u}_1 = (1, 0, 0, 0), \quad \mathbf{u}_2 = \left(0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right), \quad \mathbf{u}_3 = \left(0, \frac{1}{\sqrt{2}}, 0, -\frac{1}{\sqrt{2}}\right).$$

Then $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is ON because $\|\mathbf{u}_1\| = \|\mathbf{u}_2\| = \|\mathbf{u}_3\| = 1$ and $\mathbf{u}_1 \cdot \mathbf{u}_2 = \mathbf{u}_1 \cdot \mathbf{u}_3 = \mathbf{u}_2 \cdot \mathbf{u}_3 = 0$. ▮

Closure. In this section we introduce a dot product $\mathbf{u} \cdot \mathbf{v}$, a norm $\|\mathbf{u}\|$, and an angle θ between two vectors for n-space. Their introduction is not a matter of derivation but, rather, a matter of definition. The definitions are designed as extensions of the definitions for the familiar "arrow" vectors of 2- and 3-space, somewhat as the upper floors of a home are built upon the foundation rather than being placed on an adjacent lot. Those extensions become apparent once we express the dot product, norm, and angle for arrow vectors in n-tuple notation. The key is the definition

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n,$$

because then the arrow vector formula $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ gives us a norm, and the arrow vector formula $\theta = \cos^{-1}(\mathbf{u} \cdot \mathbf{v} / \|\mathbf{u}\| \, \|\mathbf{v}\|)$ gives us an angle θ between $\mathbf{u}$ and $\mathbf{v}$.

From these definitions we then derive the properties (12a,b,c) of the dot product and (17a,b,c) of the norm. In 2- and 3-space the triangle inequality amounts to the familiar Euclidean proposition that the length of any one side of a triangle cannot exceed the sum of the lengths of the other two sides, but in n-space, for $n > 3$, it amounts to an abstract generalization of that notion and does not have such a realizable physical or geometrical interpretation.

EXERCISES 9.5

1. Given the following vectors $\mathbf{u}$ and $\mathbf{v}$, determine $\|\mathbf{u}\|$, $\|\mathbf{v}\|$ and θ (in radians and degrees). If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, state that.

(a) $\mathbf{u} = (4, 3)$, $\mathbf{v} = (2, -1)$
(b) $\mathbf{u} = (1, 2, 3, 4)$, $\mathbf{v} = (-4, -3, -2, -1)$
(c) $\mathbf{u} = (3, 0, 1)$, $\mathbf{v} = (-2, 3, 6)$
(d) $\mathbf{u} = (2, 2, 2)$, $\mathbf{v} = (-4, -5, -6)$
(e) $\mathbf{u} = (2, 5)$, $\mathbf{v} = (10, -4)$
(f) $\mathbf{u} = (1, 2, 3, 4)$, $\mathbf{v} = (4, 3, 2, 1)$
(g) $\mathbf{u} = (3, 2, 0, -1, 1)$, $\mathbf{v} = (-5, 0, 0, 2, 4)$

2. State whether or not each of the following expressions is defined.

(a) $\|\mathbf{u}\| \, \mathbf{u}$ (b) $\mathbf{u} \cdot (\mathbf{v} \cdot \mathbf{w})$
(c) $\|(\mathbf{u} \cdot \mathbf{v})\mathbf{v}\|$ (d) $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w}$
(e) $(\mathbf{u} + \mathbf{v}) \cdot (\mathbf{u} - \mathbf{v})$ (f) $\mathbf{u} + 6(\mathbf{v} \cdot \mathbf{w})$
(g) $\cos^{-1}(2\mathbf{u} + \mathbf{v})$ (h) $\mathbf{u} / \|\mathbf{u}\|^2$
(i) $(7\mathbf{u}) \cdot (2\mathbf{v})$ (j) $\|\mathbf{u} + 3\mathbf{u}^2\|$

3. Let us denote, as points in 2- and 3-space, $A = (2, 0)$, $B = (3, -1)$, $C = (5, 0)$, $D = (4, 2)$, $E = (2, 2)$, $F = (1, 3, -2)$, $G = (2, 0, 4)$, $H = (5, 4, 3)$, $I = (-3, -1, 0)$, $J = (0, 0, 0)$. Determine, by vector methods, all interior angles and their sum, in degrees, for each of the following polygons.

(a) $ABCA$ (b) $ABCDA$ (c) $ABCDEA$
(d) $BCDB$ (e) $BCDEB$ (f) $FGHF$
(g) $FGIF$ (h) $GHIG$ (i) $FGJF$
(j) $GHJG$ (k) $HIJH$ (l) $FIJF$

4. (a)–(g) Normalize each pair of $\mathbf{u}$, $\mathbf{v}$ vectors in Exercise 1; that is, obtain $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$.

5. If vectors $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, represented as arrows, form a triangle such that $\mathbf{A} = \mathbf{B} + \mathbf{C}$, derive the *law of cosines* $C^2 = A^2 + B^2 - 2AB\cos\alpha$, where α is the interior angle between $\mathbf{A}$ and $\mathbf{B}$, and where A, B, C are the lengths of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, respectively, by starting with the identity $\mathbf{C} \cdot \mathbf{C} = (\mathbf{A} - \mathbf{B}) \cdot (\mathbf{A} - \mathbf{B})$.

6. (*Orthogonalization*) In each of following, find scalars α, β, γ and vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ such that $\mathbf{u}_1 = \mathbf{u}$, $\mathbf{u}_2 = \mathbf{u} + \alpha\mathbf{v}$, $\mathbf{u}_3 = \mathbf{u} + \beta\mathbf{v} + \gamma\mathbf{w}$ is a nonzero orthogonal set, that is,

$\mathbf{u}_1 \cdot \mathbf{u}_2 = 0$, $\mathbf{u}_1 \cdot \mathbf{u}_3 = 0$, and $\mathbf{u}_2 \cdot \mathbf{u}_3 = 0$, where $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3 \neq \mathbf{0}$. If this is *not* possible, state that.

(a) $\mathbf{u} = (1,3,0)$, $\mathbf{v} = (2,3,0)$, $\mathbf{w} = (2,1,-3)$
(b) $\mathbf{u} = (2,0,-1)$, $\mathbf{v} = (1,2,3)$, $\mathbf{w} = (3,-2,-5)$
(c) $\mathbf{u} = (1,0,0)$, $\mathbf{v} = (2,1,0)$, $\mathbf{w} = (3,2,-1)$
(d) $\mathbf{u} = (1,2,0,1)$, $\mathbf{v} = (1,0,1,1)$, $\mathbf{w} = (2,-1,1,1)$
(e) $\mathbf{u} = (1,0,0,0)$, $\mathbf{v} = (1,1,0,0)$, $\mathbf{w} = (1,1,1,0)$
(f) $\mathbf{u} = (1,-1,1,-1)$, $\mathbf{v} = (1,2,0,1)$, $\mathbf{w} = (0,2,1,0)$
(g) $\mathbf{u} = (1,2)$, $\mathbf{v} = (0,2)$, $\mathbf{w} = (1,-1)$
(h) $\mathbf{u} = (3,0)$, $\mathbf{v} = (1,1)$, $\mathbf{w} = (-1,2)$

7. If $\mathbf{u} = (1,3,-4,2)$ and $\mathbf{v} = (2,0,0,3)$, evaluate the following.

(a) $\|\mathbf{u}-\mathbf{v}\|$ (b) $\|3\mathbf{u}-2\mathbf{v}\| + \|-\mathbf{v}\|$
(c) $\left\|\dfrac{1}{\|\mathbf{u}\|}\mathbf{u}\right\|$ (d) $\|\mathbf{u}\| + \|\mathbf{v}\|$

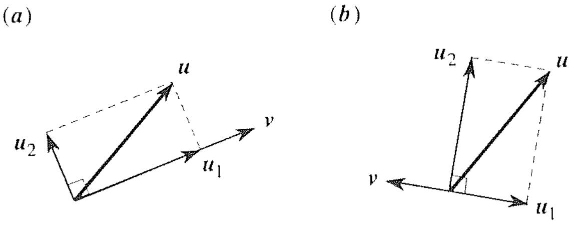
8. Derive the following identities.

(a) $\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$
(b) $\|\mathbf{u}+\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2 = 4\mathbf{u}\cdot\mathbf{v}$
(c) Verify parts (a) and (b) for the case $\mathbf{u} = (2,0,1,1)$ and $\mathbf{v} = (1,-3,0,2)$.

9. Find all nonzero vectors (if any) orthogonal to the following vectors.

(a) $(3,0,-1)$
(b) $(2,1,1)$ and $(1,2,3)$
(c) $(1,1,0,-1)$
(d) $(1,3,4,0)$ and $(2,-1,0,5)$
(e) $(6,-1,2,2)$, $(1,4,3,0)$, and $(4,-9,-4,2)$
(f) $(6,-1,2,2)$, $(1,4,3,0)$, and $(4,-5,-4,2)$
(g) $(1,-2,0)$, $(2,3,1)$, and $(7,0,2)$
(h) $(2,1,-1)$, $(1,1,1)$, and $(3,2,1)$

10. (*Orthogonal separation*) It is sometimes desired to separate a given nonzero vector $\mathbf{u}$ into the sum of two orthogonal vectors, one parallel to and one perpendicular to some other nonzero vector $\mathbf{v}$, as sketched in part (a) of the accompanying figure. That is, $\mathbf{u} = \mathbf{u}_1 + \mathbf{u}_2$, where $\mathbf{u}_1$ is of the form $\alpha\mathbf{v}$, and $\mathbf{u}_2 \cdot \mathbf{u}_1 = 0$. We call $\mathbf{u}_1$ the *orthogonal projection of* $\mathbf{u}$ *on* $\mathbf{v}$, and we call $\mathbf{u}_2$ the *component of* $\mathbf{u}$ *orthogonal to* $\mathbf{v}$.



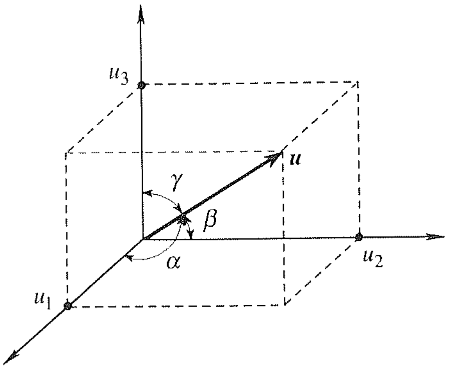
(a) Show that $\mathbf{u}_1$ and $\mathbf{u}_2$ can be found, in terms of $\mathbf{u}$ and $\mathbf{v}$, as

$$\begin{aligned} \mathbf{u}_1 &= (\mathbf{u}\cdot\hat{\mathbf{v}})\hat{\mathbf{v}} \qquad \text{where} \quad \hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|, \\ \mathbf{u}_2 &= \mathbf{u} - \mathbf{u}_1. \end{aligned} \tag{10.1}$$

Is (10.1) valid only for 2- and 3-space, or does it hold, without modification, for n-space as well? Explain.
(b) Use (10.1) to carry out the separation for the cases where $\mathbf{u} = (1,3)$ and $\mathbf{v} = (2,4)$, and where $\mathbf{u} = (1,3)$ and $\mathbf{v} = (-1,-2)$. Interpret your results graphically for each of these cases.
(c) Use (10.1) to carry out the separation for $\mathbf{u} = (2,3,1)$ and $\mathbf{v} = (0,2,3)$.
(d) Repeat part (c), for $\mathbf{u} = (1,2,-1)$, $\mathbf{v} = (3,-1,1)$.
(e) Repeat part (c), for $\mathbf{u} = (3,0,5,6)$, $\mathbf{v} = (1,-2,0,4)$.
(f) Repeat part (c), for $\mathbf{u} = (2,1,0,0,3)$, $\mathbf{v} = (0,0,1,-2,1)$.

11. (a) Prove the associative property $(\alpha\mathbf{u})\cdot\mathbf{v} = \alpha(\mathbf{u}\cdot\mathbf{v})$.
(b) Prove the distributive property $(\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = \mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w}$.
(c) Prove that the linearity property (12c) is equivalent to the two properties given in parts (a) and (b).

12. (*Direction cosines*) The **direction cosines** of a vector $\mathbf{u} = (u_1, u_2, u_3)$ in 3-space are defined as $l_1 \equiv \cos\alpha$, $l_2 \equiv \cos\beta$, $l_3 \equiv \cos\gamma$, where α, β, γ are the angles between $\mathbf{u}$ and the positive coordinate axes, as shown.



(a) Obtain general expressions for l_1, l_2, l_3 in terms of the components u_1, u_2, u_3.
(b) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (2,-1,5)$.
(c) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (2,-4,1)$.
(d) Evaluate l_1, l_2, l_3 for $\mathbf{u} = (4,0,-3)$.
(e) Show that $l_1^2 + l_2^2 + l_3^2 = 1$.

13. If $\mathbf{u}\cdot\mathbf{v} = 0$ and $\mathbf{v}\cdot\mathbf{w} = 0$, does that imply that $\mathbf{u}\cdot\mathbf{w} = 0$? Prove or disprove. HINT: If a claim is true, it needs to be proved in general, that is, for all possible cases. But if it is false, it can be disproved merely by putting forward a single counterexample.

14. Determine whether or not each set of vectors is orthogonal. (a) $(1,3),\ (-6,2),\ (0,0)$
(b) $(2,3,0),\ (-3,2,1),\ (1,1,1),\ (1,-3,1)$
(c) $(1,0,0,0),\ (0,1,0,0),\ (0,0,1,0),\ (0,0,0,1)$
(d) $(1,1,1,1),\ (1,-1,1,-1),\ (0,1,0,-1),\ (2,0,-2,0)$
(e) $(2,1,-1,1),\ (1,1,3,0),\ (1,-1,0,-1),\ (2,1,1,1)$

15. Determine a unit vector along the line of intersection of the following two planes in $\mathbb{R}^3$. NOTE: Do not use "cross products" since this topic has not yet been discussed here.

(a) $\begin{aligned} x_1 + 2x_2 - x_3 &= 8 \\ x_1 - x_2 + x_3 &= 0 \end{aligned}$

(b) $\begin{aligned} x_1 + x_2 \qquad &= 0 \\ x_1 - x_2 + 2x_3 &= 0 \end{aligned}$

(c) $\begin{aligned} x_1 - x_2 - 5x_3 &= 0 \\ x_2 + 4x_3 &= 6 \end{aligned}$

(d) $\begin{aligned} 2x_1 \qquad + x_3 &= 1 \\ x_1 + 4x_2 \qquad &= 1 \end{aligned}$

(e) $\begin{aligned} x_1 - 5x_2 + x_3 &= 4 \\ 2x_1 - x_2 - x_3 &= 3 \end{aligned}$

(f) $\begin{aligned} x_1 + x_2 + 12x_3 &= 0 \\ x_1 + 2x_2 + 12x_3 &= 5 \end{aligned}$

(g) $\begin{aligned} x_1 - x_2 - x_3 &= 2 \\ x_1 - x_2 - 2x_3 &= 5 \end{aligned}$

16. (*Schwarz inequality*) To make (15) as close to an equality as possible, and hence as informative as possible, we minimized the left-hand side by setting $d(\text{left-hand side})/d\alpha = 0$. That step gave $\alpha = -\mathbf{u}\cdot\mathbf{v}/\,\|\mathbf{v}\|^2$, and putting that result back into (15) gave the Schwarz inequality. That proof is valid for $\mathbb{R}^n$ for any $n\ (\geq 1)$. For the special case of $\mathbb{R}^2$, show that the optimal α is $-\mathbf{u}\cdot\mathbf{v}/\,\|\mathbf{v}\|^2$ by using a graphical approach; that is, using a suitable sketch. HINT: Given $\mathbf{u}$ and $\mathbf{v}$, make $\mathbf{u}+\alpha\mathbf{v}$ as short as possible.

9.6 Generalized Vector Space

9.6.1. Vector space. In Section 9.5 we generalize our vector concept from the familiar arrow vectors of 2- and 3-space to n-tuple vectors in abstract n-space, and it is n-space that is used in the remainder of this chapter and in Chapters 10–12. Yet, it is interesting to wonder if further generalization is possible. The answer is yes, and we will complete that story in this section. Far from being just a mathematical curiosity, the results will be essential in later chapters, when we study Fourier series, Sturm–Liouville theory, and partial differential equations.

The idea is as follows. In preceding sections we introduced the vectors and arithmetic rules for their manipulation, and *then* derived the various properties, such as $\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}$, $\mathbf{u}+\mathbf{0}=\mathbf{u}$, $\alpha(\beta\mathbf{u})=(\alpha\beta)\mathbf{u}$, and so on. In generalizing, the essential idea is to reverse the cart and the horse. Specifically, we elevate the derived properties to axioms, or requirements, and regard the vectors as "objects," the nature of which is not restricted in advance. They may be chosen to be n-tuples or whatever; all that we ask is that a plus $(+)$ operation, a zero vector, a negative inverse, and scalar multiplication be defined such that all of the vector space axioms are satisfied. Thus:

DEFINITION 9.6.1 *Vector Space*
We call a (nonempty) set S of "objects," which are denoted by **boldface type** and referred to as *vectors*, a **vector space** if the following requirements are met:

(i) An operation, which will be called vector *addition* and denoted as +, is defined between any two vectors in $\mathcal{S}$ in such a way that if $\mathbf{u}$ and $\mathbf{v}$ are in $\mathcal{S}$, then $\mathbf{u}+\mathbf{v}$ is too (i.e., $\mathcal{S}$ is *closed under addition*). Furthermore,

$$\mathbf{u}+\mathbf{v}=\mathbf{v}+\mathbf{u}, \qquad \text{(commutative)} \tag{1}$$

$$(\mathbf{u}+\mathbf{v})+\mathbf{w}=\mathbf{u}+(\mathbf{v}+\mathbf{w}). \qquad \text{(associative)} \tag{2}$$

(ii) $\mathcal{S}$ contains a unique *zero vector* $\mathbf{0}$ such that

$$\mathbf{u}+\mathbf{0}=\mathbf{u} \tag{3}$$

for each $\mathbf{u}$ in $\mathcal{S}$.

(iii) For each $\mathbf{u}$ in $\mathcal{S}$ there is a unique vector "$-\mathbf{u}$" in $\mathcal{S}$, called the *negative inverse of* $\mathbf{u}$, such that

$$\mathbf{u}+(-\mathbf{u})=\mathbf{0}. \tag{4}$$

We denote $\mathbf{u}+(-\mathbf{v})$ as $\mathbf{u}-\mathbf{v}$ for brevity, but emphasize that it is actually the + operation between $\mathbf{u}$ and $-\mathbf{v}$.

(iv) Another operation, called *scalar multiplication*, is defined such that if $\mathbf{u}$ is any vector in $\mathcal{S}$ and α is any scalar,* then the scalar multiple $\alpha\mathbf{u}$ is in $\mathcal{S}$, too (i.e., $\mathcal{S}$ is *closed under scalar multiplication*). Further, we require that

$$\alpha(\beta\mathbf{u})=(\alpha\beta)\mathbf{u}, \qquad \text{(associative)} \tag{5}$$

$$(\alpha+\beta)\mathbf{u}=\alpha\mathbf{u}+\beta\mathbf{u}, \qquad \text{(distributive)} \tag{6}$$

$$\alpha(\mathbf{u}+\mathbf{v})=\alpha\mathbf{u}+\alpha\mathbf{v}, \qquad \text{(distributive)} \tag{7}$$

$$1\mathbf{u}=\mathbf{u}, \tag{8}$$

if the vectors $\mathbf{u}$, $\mathbf{v}$ are in $\mathcal{S}$, and α, β are scalars.

Observe that if we write $\mathbf{u}+\mathbf{v}+\mathbf{w}$, it is not clear whether we mean $(\mathbf{u}+\mathbf{v})+\mathbf{w}$ (i.e., first add $\mathbf{u}$ and $\mathbf{v}$, and then add the result to $\mathbf{w}$) or $\mathbf{u}+(\mathbf{v}+\mathbf{w})$. However, the associative property (2) guarantees that it does not matter, so the parentheses can be omitted without ambiguity. Similarly, $\alpha\beta\mathbf{u}$ is unambiguous by virtue of (5).

EXAMPLE 1. $\mathbb{R}^n$*-Space.* Surely, the n-space $\mathbb{R}^n$, defined earlier, does constitute a vector space; after all, the axioms listed in Definition 9.6.1 come from the properties of $\mathbb{R}^n$ listed in Section 9.4. Thus, there is no need to check to see if those axioms are satisfied.

Instead, and for heuristic purposes, let us modify our addition operation from

$$\mathbf{u}+\mathbf{v}\equiv(u_1+v_1,\ldots,u_n+v_n) \tag{9}$$

*We continue to restrict all scalars to be (finite) real numbers. Hence, we call the vector space a **real vector space**.

to

$$\mathbf{u} + \mathbf{v} \equiv (u_1 + 2v_1, \dots, u_n + 2v_n), \tag{10}$$

and see if (10) works; that is, let us see if the vector space axioms listed under (i) in Definition 9.6.1 are still satisfied if we use (10) as our addition operation instead of (9). According to (10),

$$\mathbf{v} + \mathbf{u} \equiv (v_1 + 2u_1, \dots, v_n + 2u_n) \tag{11}$$

so a comparison of (10) and (11) shows that the commutativity axiom (1) is satisfied only if $u_j + 2v_j = v_j + 2u_j$ $(j = 1, \dots, n)$, hence only if $v_j = u_j$, hence only if $\mathbf{v} = \mathbf{u}$. Since (1) does not hold for any chosen vectors $\mathbf{u}$ and $\mathbf{v}$, but only for vectors $\mathbf{u}$ and $\mathbf{v}$ that are equal, we conclude that if $\mathbf{u} + \mathbf{v}$ is defined by (10), then we do *not* have a vector space. Of course, it is possible that (10) violates other axioms besides (1), but one failure is sufficient to show that the set is not a legitimate vector space.

COMMENT. Observe that we have not shown that $\mathbf{u} + \mathbf{v}$ *must* be defined as in (9); conceivably,

$$\mathbf{u} + \mathbf{v} \equiv (u_1^2 + v_1, \dots, u_n^2 + v_n) \tag{12}$$

or

$$\mathbf{u} + \mathbf{v} \equiv (u_1 - v_1, \dots, u_n - v_n) \tag{13}$$

might work; that is, might satisfy the requirements listed under (i). Thus, understand that the plus signs on the left- and right-hand sides of (9) are not the same. The ones on the right denote the usual addition of real numbers (e.g., $2 + 5 = 7$), whereas the one on the left is more exotic; it denotes a certain operation between vectors $\mathbf{u}$ and $\mathbf{v}$, which is being defined by (9), or (10), or (12), or (13). To emphasize that point we could use a different notation such as $\mathbf{u} * \mathbf{v}$, in place of $\mathbf{u} + \mathbf{v}$, as some authors do. However, having made that point let us continue to use $\mathbf{u} + \mathbf{v}$. ▮

$\mathbb{R}^n$ is but one example of a vector space. Many other useful spaces can be introduced by using objects other than n-tuples as the vectors. For example, the vectors may be functions, matrices, or whatever, provided that vector addition, a zero vector, a negative inverse, and scalar multiplication are defined such that all of the vector space axioms are satisfied. For nowhere in Definition 9.6.1 is the *nature* of the vectors specified or in any way restricted.

EXAMPLE 2. *A Function Space.* This time, let the vectors be functions. Specifically, let $\mathbf{u} = u(x)$ be any continuous function defined on $0 \leq x \leq 1$, say. For the addition operation let

$$\mathbf{u} + \mathbf{v} \equiv u(x) + v(x); \tag{14a}$$

that is, let $\mathbf{u} + \mathbf{v}$ be the function whose values are the ordinary sum $u(x) + v(x)$. For scalar multiplication let

$$\alpha\mathbf{u} \equiv \alpha u(x); \tag{14b}$$

for the zero vector choose the zero function

$$\mathbf{0} \equiv 0; \tag{14c}$$

and for the negative of $\mathbf{u}$ define

$$-\mathbf{u} \equiv -u(x); \tag{14d}$$

that is, the function whose values are $-u(x)$.

With these definitions, we can verify that all of the vector space requirements are satisf.ed, so that the set $\mathcal{S}$ of such vectors is a bona fide vector space. For instance, if $\mathbf{u} = u(x)$ and $\mathbf{v} = v(x)$ are continuous on $0 \leq x \leq 1$, then so is $\mathbf{u} + \mathbf{v} = u(x) + v(x)$ so $\mathcal{S}$ is closed under addition. Further, $\mathbf{v} + \mathbf{u} = v(x) + u(x) = u(x) + v(x) = \mathbf{u} + \mathbf{v}$,* so addition satisfies the commutative property (1), and so on.

This $\mathcal{S}$ is but one example of a **function space**, a space in which the vectors are functions. ▮

The following theorem is useful, and its proof illustrates the axiomatic approach.

THEOREM 9.6.1 *Properties of Scalar Multiplication*
If $\mathbf{u}$ is any vector in a vector space $\mathcal{S}$ and α is any scalar, then

$$0\mathbf{u} = \mathbf{0}, \tag{15a}$$

$$(-1)\mathbf{u} = -\mathbf{u}, \tag{15b}$$

$$\alpha\mathbf{0} = \mathbf{0}. \tag{15c}$$

Proof: These results follow from our definition of vector space. To prove (15a), one line of approach is as follows:

$$\begin{aligned} 0\mathbf{u} + \mathbf{u} &= 0\mathbf{u} + 1\mathbf{u} && \text{by (8)} \\ &= (0+1)\mathbf{u} && \text{by (6)} \\ &= 1\mathbf{u} \\ &= \mathbf{u} && \text{by (8).} \end{aligned}$$

Then

$$\begin{aligned} 0\mathbf{u} + \mathbf{u} + (-\mathbf{u}) &= \mathbf{u} + (-\mathbf{u}) \\ 0\mathbf{u} + \mathbf{0} &= \mathbf{0} && \text{by (4),} \\ 0\mathbf{u} &= \mathbf{0} && \text{by (3).} \end{aligned}$$

The remaining two, (15b) and (15c), are left for the exercises. ∎

9.6.2. Inclusion of inner product and/or norm. Observe that there is no mention of a dot product or a norm either in Definition 9.6.1 or in Examples 1 or 2. Indeed, a vector space $\mathcal{S}$ need not *have* a dot product (also called an **inner product**) or a norm defined for it. If it does have an inner product it is called an *inner product*

*The second equality holds because $v(x) + u(x)$ is the ordinary sum of two real numbers; e.g., $4 + 3 = 3 + 4$.

space; if it has a norm it is called a *normed vector space*; and if it has both it is called a *normed inner product space*.

If we do choose to introduce an inner product for $\mathcal{S}$, how is it to be defined? Do you remember the idea of reversing the cart and the horse? That is how we do it. Equations (12a,b,c) in Section 9.5.2 were shown to be properties of the inner product $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$. We now take those properties and elevate them to axioms, or requirements, that are to be satisfied by any inner product of any vector space.

Similarly, we take the properties (17a,b,c) of the norm, in Section 9.5.3, and elevate them to axioms, or requirements, that are to be satisfied by any norm of any vector space.

Let us tabulate them here:

REQUIREMENTS OF INNER PRODUCT

Commutative: $$\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}, \tag{16a}$$

Nonnegative: $$\begin{aligned} \mathbf{u} \cdot \mathbf{u} &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, \end{aligned} \tag{16b}$$

Linear: $$(\alpha\mathbf{u} + \beta\mathbf{v}) \cdot \mathbf{w} = \alpha(\mathbf{u} \cdot \mathbf{w}) + \beta(\mathbf{v} \cdot \mathbf{w}), \tag{16c}$$

and

REQUIREMENTS OF NORM

Scaling: $$\|\alpha\mathbf{u}\| = |\alpha| \, \|\mathbf{u}\|, \tag{17a}$$

Nonnegative: $$\begin{aligned} \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, \end{aligned} \tag{17b}$$

Triangle Inequality: $$\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|. \tag{17c}$$

Let us illustrate.

EXAMPLE 3. $\mathbb{R}^n$*-Space.* If we wish to add an inner product to the vector space $\mathbb{R}^n$, we can use the choice

$$\mathbf{u} \cdot \mathbf{v} \equiv u_1 v_1 + \cdots + u_n v_n = \sum_{j=1}^{n} u_j v_j. \tag{18a}$$

We know that (18a) satisfies the requirements (16) because the latter were deduced, in Section 9.5.2, as properties that *follow* from (18a). A variation of (18a) that still satisfies (16) is (Exercise 6)

$$\mathbf{u}\cdot\mathbf{v} \equiv w_1u_1v_1 + \cdots + w_nu_nv_n = \sum_{j=1}^{n} w_ju_jv_j, \tag{18b}$$

where the w_j's are fixed positive constants known as "weights" because they attach more or less weight to the different components of $\mathbf{u}$ and $\mathbf{v}$. For instance, consider $\mathbb{R}^2$ and let $w_1 = 5$ and $w_2 = 3$. Then if $\mathbf{u} = (2,-4)$ and $\mathbf{v} = (1,6)$ we have $\mathbf{u}\cdot\mathbf{v} = 5(2)(1) + 3(-4)(6) = -62$.

Note that for (18b) to be a legitimate inner product we must have $w_j > 0$ for each j. For suppose, still in $\mathbb{R}^2$, that $w_1 = 3$ and $w_2 = -2$. Then, for $\mathbf{u} = (1,5)$, say, we have $\mathbf{u}\cdot\mathbf{u} = 3(1)(1) - 2(5)(5) = -47 < 0$, in violation of (16b). Or, suppose that $w_1 = 3$ and $w_2 = 0$. Then, for $\mathbf{u} = (0,4)$, say, we have $\mathbf{u}\cdot\mathbf{u} = 3(0)(0) + 0(4)(4) = 0$ even though $\mathbf{u} \neq \mathbf{0}$, again in violation of (16b).

Now, suppose that we wish to add a norm. If for any vector space $\mathcal{S}$ we already have an inner product, then a legitimate norm can always be obtained from that inner product as $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$, and that choice is called the *natural norm.* Thus, the natural norms corresponding to (18a) and (18b) are

$$\|\mathbf{u}\| \equiv \sqrt{\sum_{1}^{n} u_j^2} \quad \text{and} \quad \|\mathbf{u}\| \equiv \sqrt{\sum_{1}^{n} w_ju_j^2}, \tag{19a,b}$$

respectively.

However, we do not *have* to choose the natural norm. For instance, we could use (18a) as our inner product, and choose

$$\|\mathbf{u}\| \equiv |u_1| + \cdots + |u_n| = \sum_{j=1}^{n} |u_j| \tag{20}$$

as our norm (Exercise 8). The latter is used by Struble in his book on differential equations,* probably because it is algebraically simpler than the *Euclidean norm* (19a) or the *modified Euclidean norm* (19b). Furthermore, he defines no inner product whatsoever. Struble calls (20) the *taxicab norm* since a taxicab driver judges the distance from the corner of 5th Avenue and 34th Street to the corner of 2nd Avenue and 49th Street as 18 blocks, not $\sqrt{234}$ blocks. ▮

EXAMPLE 4. *The Function Space of Example 2.* How might we choose an inner product for the *function* space $\mathcal{S}$ defined in Example 2? To motivate our choice, let us imagine approximating any given function (i.e., vector) $u(x)$ in $\mathcal{S}$ in a piecewise-constant manner as depicted in Fig. 1. That is, divide the x interval $(0 \le x \le 1)$ into n equal parts and define the approximating piecewise-constant function, over each subinterval as

*R. A. Struble, *Nonlinear Differential Equations* (New York: McGraw-Hill, 1962).

the value of $u(x)$ at the left endpoint of that subinterval. If we represent the piecewise-constant function as the n-tuple $(u_1, \ldots, u_n)$, then we have, in a heuristic sense,

$$u(x) \approx (u_1, \ldots, u_n). \tag{21}$$

Similarly, for any other function $v(x)$ in S,

$$v(x) \approx (v_1, \ldots, v_n). \tag{22}$$

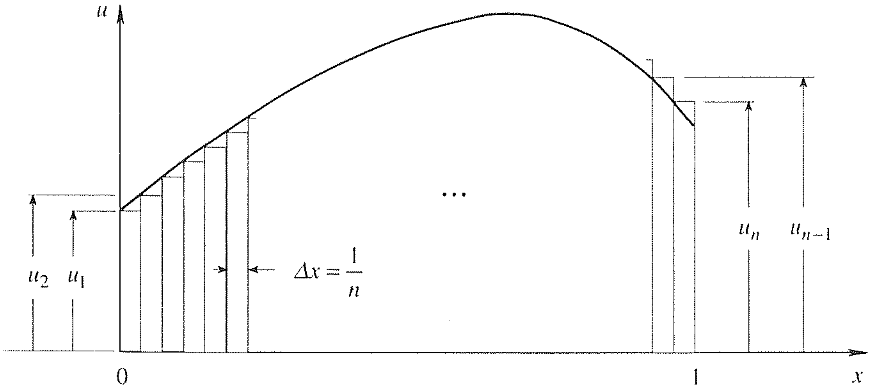


Figure 1. Staircase approximation of $u(x)$.

The n-tuple vectors on the right-hand sides of (21) and (22) are members of $\mathbb{R}^n$. For that space, let us adopt the inner product

$$(u_1, \ldots, u_n) \cdot (v_1, \ldots, v_n) = \sum_{j=1}^{n} u_j v_j \Delta x, \tag{23}$$

that is, (18b) with all of the w_j weights the same, namely, the subinterval width Δx. If we let $n \to \infty$, the "staircase approximations" approach $u(x)$ and $v(x)$, and the sum in (23) tends to the integral $\int_0^1 u(x)v(x)\,dx$.

This heuristic reasoning suggests the inner product

$$\boxed{\left\langle u(x), v(x) \right\rangle \equiv \int_0^1 u(x)v(x)\,dx.} \tag{24a}$$

We can denote it as $\mathbf{u} \cdot \mathbf{v}$ and call it the dot product, or we can denote it as $< u(x), v(x) >$ and call it the inner product. For function spaces, the latter notation is somewhat standard, and is our choice in this text.

COMMENT 1. By no means do we claim our staircase idea to be a rigorous derivation of (24a). In fact, it is neither rigorous nor a derivation; it is *heuristic motivation* for the *definition* (24a). We leave it for the exercises to verify that (24a) does satisfy the requirements (16).

COMMENT 2. Just as (18b) is a legitimate generalization of (18a), (if $w_j > 0$ for $1 \leq j \leq n$), we expect that

$$\boxed{\left\langle u(x), v(x) \right\rangle \equiv \int_0^1 u(x)v(x)w(x)\,dx} \tag{24b}$$

is a legitimate generalization of (24a) [if $w(x) > 0$ for $0 \leq x \leq 1$], proof of which claim is left for the exercises. The inner product (24b) is prominent when we study Fourier series and the Sturm–Liouville theory in Chapter 17.

COMMENT 3. Naturally, if we wish to define a norm as well, we could use a natural norm based on (24a) or (24b), for instance

$$\|\mathbf{u}\| \equiv \sqrt{< u(x), u(x) >} = \sqrt{\int_0^1 u^2(x) w(x)\, dx} \tag{25}$$

based on (24b).

COMMENT 4. Notice carefully that the concept of the *dimension* of a vector space has not yet been introduced, although it is in Section 9.10. There, we define dimension and find that $\mathbb{R}^n$ is n-dimensional (which claim is probably not a great shock). Since the staircase approximation (21) becomes exact only as $n \to \infty$, it appears that our function space S is infinite dimensional!

COMMENT 5. A bit of notation: the set of functions that are defined and continuous on $[0, 1]$ (i.e., $0 \leq x \leq 1$) is usually denoted as $C^0[0, 1]$. If not only are the functions continuous but also all derivatives through order k, then the set is denoted as $C^k[0, 1]$. ∎

Closure. Using n-space as a ladder, we complete our generalization of vector space by taking the properties of $\mathbb{R}^n$ (such as $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$) and turning them into the axioms, or requirements, to be met by *any* vector space. Thus, attention shifted from the objects, the vectors, to those requirements. There is no restriction on the nature of the vectors, which can be arrows, n-tuples, matrices, functions, or oranges. For us, the most important vector spaces are $\mathbb{R}^n$ and various function spaces; $\mathbb{R}^n$ is used in the remainder of this chapter and Chapters 10–12, and function spaces are used in Chapter 17 when we study Fourier series and Sturm-Liouville theory.

To illustrate the power of the axiomatic approach, recall the Schwarz inequality $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|\, \|\mathbf{v}\|$, proved in Section 9.5.2 for $\mathbb{R}^n$. That result holds for *any* normed inner product space with natural norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$ for it followed from properties of $\mathbb{R}^n$, which properties are subsequently elevated to axioms for general vector space. Thus, it represents many properties rolled into one. For example, in $\mathbb{R}^n$, with the dot product (18a) it says

$$\left| \sum_{j=1}^{n} u_j v_j \right| \leq \sqrt{\sum_1^n u_j^2} \sqrt{\sum_1^n v_j^2}, \tag{26}$$

in the function space of Examples 2 and 4; with the inner product (24b) and norm (25) it says

$$\left| \int_0^1 u(x) v(x) w(x)\, dx \right| \leq \sqrt{\int_0^1 u^2(x) w(x)\, dx} \sqrt{\int_0^1 v^2(x) w(x)\, dx}, \tag{27}$$

and so on.

EXERCISES 9.6

1. Recall that $\mathbb{R}^n$ is the vector space ("real" vector space since all scalars are to be real numbers) in which the vectors are n-tuples $\mathbf{u} = (u_1, \dots, u_n)$, with the definitions

$$\begin{aligned} \mathbf{u}+\mathbf{v} &= (u_1, \dots, u_n) + (v_1, \dots, v_n) \\ &\equiv (u_1+v_1, \dots, u_n+v_n), & (1.1) \\ \mathbf{0} &\equiv (0, \dots, 0), & (1.2) \\ -\mathbf{u} &\equiv (-u_1, \dots, -u_n), & (1.3) \\ \alpha\mathbf{u} &\equiv (\alpha u_1, \dots, \alpha u_n). & (1.4) \end{aligned}$$

If we make the following modifications, do we still have a vector space? If not, specify all requirements within Definition 9.6.1 that fail to be met.

(a) only vectors of the form $\mathbf{u} = (u, u, \dots, u)$ admitted, where $-\infty < u < \infty$
(b) only vectors of the form $\mathbf{u} = (u, 2u, 3u, \dots, nu)$ admitted, where $-\infty < u < \infty$
(c) only the vector $(0, \dots, 0)$ admitted (this is an example of a **zero vector space**, a vector space containing only the zero vector)
(d) $\mathbf{u}+\mathbf{v} \equiv (u_1 - v_1, \dots, u_n - v_n)$, in place of (1.1)
(e) $\mathbf{u}+\mathbf{v} \equiv (0, \dots, 0)$ for all $\mathbf{u}$'s and $\mathbf{v}$'s, in place of (1.1)
(f) $\alpha\mathbf{u} \equiv (\alpha^2 u_1, \dots, \alpha^2 u_n)$, in place of (1.4)

2. We noted in Example 1 that the definition (10) of vector addition violates axiom (1). Does it violate any others as well? Explain.

3. Prove (15b), that $(-1)\mathbf{u} = -\mathbf{u}$.

4. Prove (15c), that $\alpha\mathbf{0} = \mathbf{0}$.

5. Prove that if $\alpha\mathbf{u} = \mathbf{0}$ then $\alpha = 0$ and/or $\mathbf{u} = \mathbf{0}$.

6. Show that the inner product (18b) does satisfy the requirements (16).

7. We stated in Example 3 that if for any vector space S we already have an inner product, then a legitimate norm can always be obtained from that inner product as $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$, which choice is called the natural norm. Prove that claim.

8. Show that the "taxicab norm" (20) is a legitimate norm – that is, that it satisfies the requirements (17).

9. (a) Does the choice $\|\mathbf{u}\| = \max\limits_{1\le j\le n} |u_j|$, for $\mathbb{R}^n$, satisfy the requirements (17)? Explain.
(b) How about $\|\mathbf{u}\| = \min\limits_{1\le j\le n} |u_j|$, for $\mathbb{R}^n$?

10. Let S be the set of real-valued polynomial functions, of degree n, defined on $a \le x \le b$. If $\mathbf{u} = a_0 + a_1x + \cdots + a_nx^n$ and $\mathbf{v} = b_0 + b_1x + \cdots + b_nx^n$ are any two such functions, and α is any (real) scalar, define the sum $\mathbf{u}+\mathbf{v}$ and the scalar multiple $\alpha\mathbf{u}$ as

$$\begin{aligned} (\mathbf{u}+\mathbf{v})(x) &= (a_0+b_0) + (a_1+b_1)x + \cdots + (a_n+b_n)x^n, \\ (\alpha\mathbf{u})(x) &= \alpha a_0 + \alpha a_1 x + \cdots + \alpha a_n x^n, \end{aligned}$$

respectively. Further, let $\mathbf{0}$ be the function $0 + 0x + \cdots + 0x^n$, and let $-\mathbf{u}$ be the function $-a_0 - a_1x + \cdots - a_nx^n$. Show that S is a vector space.

11. Show that the inner product (24b) does satisfy the requirements (16).

12. (*Schwarz inequality*) We derive the Schwarz inequality

$$|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\| \qquad (12.1)$$

for $\mathbb{R}^n$ space in Section 9.5.2. The latter holds not only for $\mathbb{R}^n$ but for *any* normed inner product space with the natural norm $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$. In this exercise we simply ask you to verify (12.1) by working out the left- and right-hand sides for these specific cases:

(a) $\mathbf{u} = (3, 1, -1, 0)$ and $\mathbf{v} = (1, 2, 5, -4)$ in $\mathbb{R}^4$, with the inner product (18a)
(b) $\mathbf{u} = (1, 2, 4, -3)$ and $\mathbf{v} = (0, 4, 1, 1)$ in $\mathbb{R}^4$, with $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + 5u_2v_2 + 3u_3v_3 + 2u_4v_4$
(c) $\mathbf{u} = (1, 1, 1, 1, 1)$ and $\mathbf{v} = (2, 2, 2, 2, 2)$ in $\mathbb{R}^5$, with $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + 2u_2v_2 + 3u_3v_3 + 4u_4v_4 + 5u_5v_5$
(d) $\mathbf{u} = 2 + x$ and $\mathbf{v} = 3x^2$ in the function space of Example 4, with the inner product $\mathbf{u}\cdot\mathbf{v} = \langle u(x), v(x)\rangle = \int_0^1 u(x)v(x)\,dx$
(e) Same as (d), but with $\langle u(x), v(x)\rangle = \int_0^1 u(x)v(x)(2 + 5x)\,dx$

13. (*Solution space*) (a) Consider a set of m linear homogeneous algebraic equations in the n unknowns $x_1, \dots, x_n$, and denote each solution of the system as an n-tuple vector $\mathbf{x} = (x_1, \dots, x_n)$ in $\mathbb{R}^n$. Show that the set of all such vectors, with the usual definitions $[\mathbf{u}+\mathbf{v} = (u_1+v_1, \dots, u_n+v_n), \alpha\mathbf{u} = (\alpha u_1, \dots, \alpha u_n), -\mathbf{u} = (-u_1, \dots, -u_n), \mathbf{0} = (0, \dots, 0)]$, is a vector space. That space is called the **solution space** of the system.

(b) If the system is *non*homogeneous, is the set of solutions still a vector space? Explain.

14. (*Solution space*) Show that the solutions of a linear homogeneous differential equation (with the same definitions of $\mathbf{u}+\mathbf{v}$, $\alpha\mathbf{u}$, $-\mathbf{u}$, and $\mathbf{0}$ as in Example 2) constitute a vector space, the so-called **solution space** of that differential equation.

9.7 Span and Subspace

Here, we begin a sequence of closely related ideas: span, linear dependence, basis, expansion, and dimension. The concepts, definitions, and theorems hold for any vector space, but our illustrative examples are restricted to the n-space $\mathbb{R}^n$, this being the case of most interest in Chapters 9–12.

We begin with the idea of the "span" of a set of vectors.

DEFINITION 9.7.1 *Span*
If $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are vectors in a vector space S, then the set of all linear combinations of these vectors, that is, all vectors of the form

$$\mathbf{u} = \alpha_1\mathbf{u}_1 + \cdots + \alpha_k\mathbf{u}_k, \tag{1}$$

where $\alpha_1, \ldots, \alpha_k$ are scalars is called the **span** of $\mathbf{u}_1, \ldots, \mathbf{u}_k$ and is denoted as span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.

The set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is called the **generating set** of span $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$.

Let us illustrate with some vector sets in $\mathbb{R}^2$ and $\mathbb{R}^3$ so we can support the discussion with diagrams.

EXAMPLE 1. Determine the span of the single vector

$$\mathbf{u}_1 = (4, 2) \tag{2}$$

in $\mathbb{R}^2$. Then span $\{\mathbf{u}_1\}$ is the set of all vectors that are scalar multiples of $\mathbf{u}_1$. Hence, span $\{\mathbf{u}_1\}$ is the set of all vectors on the line L in Fig. 1, such as $\mathbf{u} = 2\mathbf{u}_1 = (8, 4)$, $\mathbf{v} = -\frac{1}{2}\mathbf{u}_1 = (-2, -1)$, and $\mathbf{0} = 0\mathbf{u}_1 = (0, 0)$. We say that $\mathbf{u}_1$ *generates* the line L. ∎

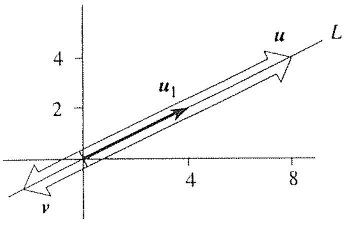


Figure 1. Span $\{\mathbf{u}_1\}$.

EXAMPLE 2. Determine the span of the two vectors

$$\mathbf{u}_1 = (4, 2), \quad \mathbf{u}_2 = (-8, -4). \tag{3}$$

Span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is, once again, the line L in Fig. 1 (i.e., the set of all vectors on L), for both $\mathbf{u}_1$ and $\mathbf{u}_2$ lie along L, so any linear combination of them, $\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2$, does too. Similarly, span $\{(4,2),\ (-8,-4),\ (18,9),\ (0,0)\}$ is the line L. ∎

Observe that the line L, in Examples 1 and 2, is only a subset of the vector space $\mathbb{R}^2$. Observe that that subset of $\mathbb{R}^2$ is itself a vector space, a so-called "subspace" of $\mathbb{R}^2$. For if $\mathbf{u}$ and $\mathbf{v}$ are any two vectors on L, then $\mathbf{u}+\mathbf{v}$ is on L, too, so the set *is* closed under addition; similarly, if $\mathbf{u}$ is on L, so is $\alpha\mathbf{u}$, for any scalar α, so the set *is* closed under scalar multiplication; L *does* contain the zero vector [since we can set all the α's in (1) equal to zero]; and for each $\mathbf{u}$ on L there is a (unique) vector $-\mathbf{u}$ on L such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.

DEFINITION 9.7.2 *Subspace*
If a subset $\mathcal{T}$ of a vector space $\mathcal{S}$ is itself a vector space (with the same definitions as $\mathcal{S}$ for vector addition $\mathbf{u}+\mathbf{v}$, scalar multiplication $\alpha\mathbf{u}$, zero vector $\mathbf{0}$, and negative vector $-\mathbf{u}$), then $\mathcal{T}$ is a **subspace** of $\mathcal{S}$.

Usually, a subspace of $\mathcal{S}$ is only a part of $\mathcal{S}$, as the line L is only a part of $\mathbb{R}^2$, but since a subset of a set can be all of that set, a subspace of $\mathcal{S}$ can be all of $\mathcal{S}$. For instance, $\mathbb{R}^2$ is a subspace of $\mathbb{R}^2$.

THEOREM 9.7.1 *Span as Subspace*
If $\mathbf{u}_1, \dots, \mathbf{u}_k$ are vectors in a vector space $\mathcal{S}$, then span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is itself a vector space, a subspace of $\mathcal{S}$.

For instance, the line L in Fig. 1 is a subspace of $\mathbb{R}^2$. Proof of Theorem 9.7.1 is left for the exercises.

EXAMPLE 3. Is the span of

$$\mathbf{u}_1 = (5,1), \quad \mathbf{u}_2 = (1,3) \tag{4}$$

all of $\mathbb{R}^2$ or only a part of $\mathbb{R}^2$? To determine the extent of span $\{\mathbf{u}_1, \mathbf{u}_2\}$, let $\mathbf{v} = (v_1, v_2)$ be any given vector in $\mathbb{R}^2$, and try to express

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2. \tag{5}$$

That is,

$$\begin{aligned}(v_1, v_2) &= \alpha_1(5,1) + \alpha_2(1,3)\\ &= (5\alpha_1, \alpha_1) + (\alpha_2, 3\alpha_2).\\ &= (5\alpha_1 + \alpha_2, \alpha_1 + 3\alpha_2).\end{aligned} \tag{6}$$

Equating components, we obtain the linear equations

$$\begin{aligned} 5\alpha_1 + \ \alpha_2 &= v_1, \\ \alpha_1 + 3\alpha_2 &= v_2 \end{aligned} \tag{7}$$

in α_1, α_2. Applying Gauss elimination, (7) becomes

$$\begin{aligned} \alpha_1 + \tfrac{1}{5}\alpha_2 &= \tfrac{1}{5}v_1, \\ \alpha_2 &= \tfrac{5}{14}v_2 - \tfrac{1}{14}v_1. \end{aligned} \tag{8}$$

It is clear from the Gauss-reduced form (8) that the system is *consistent* (solvable for α_1, α_2) for *every* vector $\mathbf{v}$ in $\mathbb{R}^2$. Hence, we may conclude that span$\{\mathbf{u}_1, \mathbf{u}_2\}$ is all of $\mathbb{R}^2$; we say that $\{\mathbf{u}_1, \mathbf{u}_2\}$ spans $\mathbb{R}^2$. (Here we use "span" as a verb; in Definition 9.7.1 it is introduced as a noun.)

Thus, every $\mathbf{v}$ in $\mathbb{R}^2$ can be expressed as a linear combination of vector $\mathbf{u}_1$ and $\mathbf{u}_2$. As representative, let $\mathbf{v} = (6, -4)$ so $v_1 = 6$ and $v_2 = -4$. Then (8) gives $\alpha_2 = -\frac{13}{7}$ and $\alpha_1 = \frac{11}{7}$, so that (5) becomes

$$\mathbf{v} = \tfrac{11}{7}\mathbf{u}_1 - \tfrac{13}{7}\mathbf{u}_2. \tag{9}$$

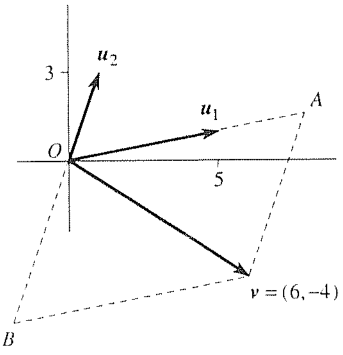


Figure 2. Representation of $\mathbf{v}$.

To see this in graphical terms, observe from Fig. 2 that $\mathbf{v} = \mathbf{OA} + \mathbf{OB}$, where (with the aid of a scale) $\mathbf{OA} \approx 1.6\mathbf{u}_1$ and $\mathbf{OB} \approx -1.9\mathbf{u}_2$. Thus, $\mathbf{v} \approx 1.6\mathbf{u}_1 - 1.9\mathbf{u}_2$, in agreement with (9).

COMMENT. Suppose that we add $\mathbf{u}_3 = (2, 2)$ to the set. It should be evident that span$\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is all of $\mathbb{R}^2$, again, since $\{\mathbf{u}_1, \mathbf{u}_2\}$ spanned $\mathbb{R}^2$ even "without any help" from $\mathbf{u}_3$. But in case this is not clear, let us go through steps analogous to steps (5) to (8):

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \alpha_3\mathbf{u}_3 \tag{10}$$

so $(v_1, v_2) = (5\alpha_1 + \alpha_2 + 2\alpha_3, \alpha_1 + 3\alpha_2 + 2\alpha_3)$. Thus,

$$\begin{aligned} 5\alpha_1 + \ \alpha_2 + 2\alpha_3 &= v_1, \\ \alpha_1 + 3\alpha_2 + 2\alpha_3 &= v_2, \end{aligned}$$

or

$$\begin{aligned} \alpha_1 + \tfrac{1}{5}\alpha_2 + \tfrac{2}{5}\alpha_3 &= \tfrac{1}{5}v_1, \\ \alpha_2 + \tfrac{4}{7}\alpha_3 &= \tfrac{5}{14}v_2 - \tfrac{1}{14}v_1. \end{aligned} \tag{11}$$

Like (8), (11) is *consistent* for every $\mathbf{v}$ in $\mathbb{R}^2$ so $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ spans $\mathbb{R}^2$, as claimed. Whereas (8) had a unique solution so that the representation (5) was unique, (11) happens to have an infinity of solutions so that the representation (10) is *not* unique. ▮

EXAMPLE 4. As a final example, consider the span of

$$\mathbf{u}_1 = (1, 2, 2), \quad \mathbf{u}_2 = (-1, 0, 2) \tag{12}$$

in $\mathbb{R}^3$. Setting

$$\mathbf{v} = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2, \tag{13}$$

we have

$$\begin{aligned}\alpha_1 - \alpha_2 &= v_1,\\ 2\alpha_1 \quad &= v_2,\\ 2\alpha_1 + 2\alpha_2 &= v_3,\end{aligned}$$

or, after Gauss elimination,

$$\begin{aligned}\alpha_1 - \alpha_2 &= v_1,\\ \alpha_2 &= \tfrac{1}{2}v_2 - v_1,\\ 0 &= v_3 - 2v_2 + 2v_1.\end{aligned} \tag{14}$$

Now, span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the set of all possible vectors $\mathbf{v}$ given by (13), i.e., all vectors $\mathbf{v}$ for which the system (14) is consistent, i.e., all vectors $\mathbf{v} = (v_1, v_2, v_3)$ such that

$$2v_1 - 2v_2 + v_3 = 0 \tag{15}$$

[so that the last of equations (14) is $0 = 0$ rather than a contradiction].

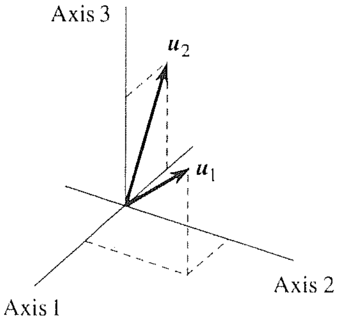


Figure 3. $\mathbf{u}_1$ and $\mathbf{u}_2$.

In geometrical terms, on the other hand, span $\{\mathbf{u}_1, \mathbf{u}_2\}$ should be the subset of $\mathbb{R}^3$ consisting of the *plane* that passes through $\mathbf{u}_1$ and $\mathbf{u}_2$ ($\mathbf{u}_1$ and $\mathbf{u}_2$ are shown in Fig. 3). How does that fact correlate with (15)? As a matter of fact, (15) *is* the equation of a plane in 3-space, and that plane does pass through the origin, through the tip of $\mathbf{u}_1$ [i.e., the point $(1, 2, 2)$], and through the tip of $\mathbf{u}_2$ [the point $(-1, 0, 2)$]. Hence, it *is* the plane through $\mathbf{u}_1$ and $\mathbf{u}_2$ so the analytical approach, namely, steps (13) to (15) and our geometrical interpretation are in agreement.

We conclude that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is not all of $\mathbb{R}^3$; it is only the subspace of $\mathbb{R}^3$ consisting of the plane (i.e., all vectors in the plane) containing the given vectors $\mathbf{u}_1$ and $\mathbf{u}_2$.

COMMENT. Since span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is a plane, would it be correct to say that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is $\mathbb{R}^2$? No, that would be incorrect; $\mathbb{R}^2$ is made up of *two*-tuples, while the vectors in the above-mentioned plane are *three*-tuples. Thus, $\mathbb{R}^2$ space is not relevant in this problem. All that can be said here is that span $\{\mathbf{u}_1, \mathbf{u}_2\}$ is the subspace of $\mathbb{R}^3$ consisting of the plane containing the vectors $\mathbf{u}_1$ and $\mathbf{u}_2$, that is, the plane defined by (15). ▮

Closure. In leading up to the concept of bases and expansions, the two key ideas are span and linear independence. In this section we introduce the idea of span; in the next section we introduce linear dependence and linear independence. Although the concept of span holds for any vector space, such as $\mathbb{R}^6$, we suggest that you focus on the foregoing examples in two- and three-spaces, so that you can use the two- and three-dimensional drawings to promote understanding.

EXERCISES 9.7

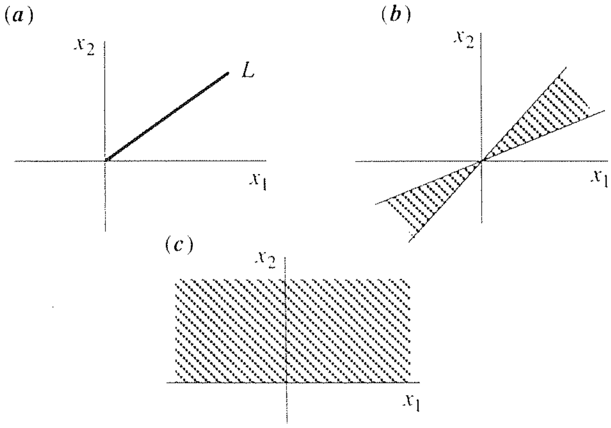
1. Show whether the vectors

(a) $(1, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0)$, $\ldots$, $(0, \ldots, 0, 1)$ span $\mathbb{R}^n$
(b) $(0, 0, 0, 1)$, $(0, 0, 1, 1)$, $(0, 1, 1, 1)$, $(1, 1, 1, 1)$ span $\mathbb{R}^4$
(c) $(1, 2, 0, 4)$, $(2, 3, 1, -1)$, $(0, 1, 0, 1)$, $(0, 0, 0, 0)$, $(1, 1, 2, 3)$ span $\mathbb{R}^4$
(d) $(1, 3, 2, 2)$, $(5, 7, 1, 0)$, $(-1, -2, -4, 3)$ span $\mathbb{R}^4$
(e) $(1, 0, 1)$, $(2, 1, -1)$, $(1, 2, -5)$ span $\mathbb{R}^3$
(f) $(1, 1, 2)$, $(0, 0, 0)$, $(2, 1, 0)$, $(-1, 0, 3)$ span $\mathbb{R}^3$
(g) $(2, 0, 3)$, $(-1, 2, 4)$, $(-5, 2, -2)$ span $\mathbb{R}^3$
(h) $(1, 3, 0)$, $(2, -1, 1)$, $(1, 1, 4)$ span $\mathbb{R}^3$
(i) $(-1, 2, 4)$, $(-5, 2, -2)$, $(2, 0, 3)$, $(1, 2, 3)$ span $\mathbb{R}^3$
(j) $(0, 0, 0)$, $(2, 1, 4)$, $(-1, 3, 5)$ span $\mathbb{R}^3$
(k) $(2, 1, 3)$, $(1, -1, 2)$ span $\mathbb{R}^3$
(l) $(2, 1, -1)$, $(1, 3, 1)$, $(5, 5, -1)$, $(0, 5, 3)$ span $\mathbb{R}^3$
(m) $(-4, 1, 0)$, $(2, 2, 2)$, $(1, 2, 3)$ span $\mathbb{R}^3$
(n) $(-3, 1, 0)$, $(1, 1, 1)$, $(-1, 7, 5)$ span $\mathbb{R}^3$
(o) $(1, 2)$, $(2, 1)$ span $\mathbb{R}^2$
(p) $(1, 2)$, $(2, 1)$, $(4, 5)$ span $\mathbb{R}^2$
(q) $(1, 2)$, $(2, 1)$, $(2, 3)$, $(2, -4)$ span $\mathbb{R}^2$

2. (a) Sketch any two vectors that span the space of all vectors in the plane of the paper.
(b) Sketch any three such vectors.
(c) Sketch any four such vectors.

3. Are the following vector sets subspaces of $\mathbb{R}^2$? (See accompanying figure.) Explain.

(a) the straight line L that extends from the origin to infinity
(b) the wedge-shaped region (including its boundary lines) that extends to infinity in both directions
(c) the upper half plane $x_2 \geq 0$



4. (*Solution space*) First, review Exercise 13a in Section 9.6. That solution space is a subspace of $\mathbb{R}^n$. To illustrate, consider the simple system $x_1 + 3x_2 = 0$; that is, $m = 1$, $n = 2$, $a_{11} = 1$, and $a_{12} = 3$. The solution is $x_2 = \alpha$ (arbitrary), $x_1 = -3\alpha$, or $\mathbf{x} = (x_1, x_2) = \alpha(-3, 1)$ so the solution space is the span of the vector $(-3, 1)$, that is, span $\{(-3, 1)\}$. In this manner, determine the solution space for each of the following examples.

(a) $x_1 - x_2 + 4x_3 = 0$ in $\mathbb{R}^3$
(b) $x_1 + x_2 + x_3 - x_4 = 0$ in $\mathbb{R}^4$
(c)
$$\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ x_1 + x_2 + x_3 &= 0 \quad \text{in } \mathbb{R}^3 \end{aligned}$$
(d)
$$\begin{aligned} x_1 + 3x_2 - x_3 + x_4 &= 0 \\ x_1 \qquad + 2x_3 + x_4 &= 0 \quad \text{in } \mathbb{R}^4 \end{aligned}$$
(e)
$$\begin{aligned} x_1 - x_2 + x_3 - 2x_4 \qquad &= 0 \\ x_1 - x_2 \qquad + x_4 + 2x_5 &= 0 \quad \text{in } \mathbb{R}^5 \end{aligned}$$
(f)
$$\begin{aligned} x_1 + x_2 - x_3 + x_4 &= 0 \\ x_1 + 2x_2 \qquad - x_4 &= 0 \\ x_2 + x_3 \qquad &= 0 \quad \text{in } \mathbb{R}^4 \end{aligned}$$
(g)
$$\begin{aligned} x_1 + x_2 + 2x_3 - 2x_4 \qquad &= 0 \\ x_1 + x_2 + 2x_3 \qquad + x_5 &= 0 \\ 2x_4 + x_5 &= 0 \quad \text{in } \mathbb{R}^5 \end{aligned}$$

5. Find any two vectors in $\mathbb{R}^3$ that span the plane

(a) $x_1 - 2x_2 + 4x_3 = 0$
(b) $2x_1 + x_2 - 6x_3 = 0$
(c) $x_1 + 5x_3 = 0$
(d) $x_1 + 4x_2 + x_3 = 0$
(e) $x_2 + 2x_3 = 0$
(f) $3x_1 - x_2 - x_3 = 0$

6. Show whether the given sets are identical. Explain.

(a) span $\{(2, -1, -1), (3, 1, 0)\}$ and span $\{(2, -1, -1), (5, 5, 2)$
(b) span $\{(1, 2, 3), (2, -1, 1)\}$ and span $\{(1, 2, 3), (3, 1, 5)\}$
(c) span $\{(4, 1, 0), (1, 1, 1)\}$ and span $\{(1, 1, 1), (2, -1, -2)\}$
(d) span $\{(1, 2, -1), (-3, 0, 0)\}$ and span $\{(1, 0, 0), (1, 3, 0)\}$
(e) span $\{(1, 0, 1, 2), (-1, 1, 1, 0)\}$ and span $\{(0, 1, 2, 2), (1, 1, 3, 4)\}$
(f) span $\{(1, 0, 1, 2), (1, 1, 1, 1), (1, 2, 3, 4)\}$ and span $\{(2, 0, -1, 0), (0, -1, 2, 3), (4, 3, 2, 1)\}$
(g) span $\{(1, 0, 1, 1), (2, 1, 1, 0), (1, 2, 2, 1)\}$ and span $\{(2, -1, 0, 0), (1, -2, 0, 1), (3, 5, 4, 1)\}$
(h) span $\{(1, 2, 3, 0), (0, 1, 0, 2), (2, 3, 0, 1)\}$ and span $\{(1, 0, -3, -1), (-1, 1, 3, 3), (1, 2, 1, 1)\}$

7. Find any two ON (orthonormal) vectors in

(a) span $\{(1, 2), (6, -1)\}$
(b) span $\{(1, 2, 4), (2, -1, 3)\}$
(c) span $\{(1, -1, 0), (1, 2, 3)\}$

(d) span $\{(2, 1, 0), (0, 1, 2)\}$
(e) span $\{(1, 1, 0, 1), (0, 2, -1, 1)\}$
(f) span $\{(-2, 3, 1, 1), (0, 2, -1, 1)\}$

8. Prove Theorem 9.7.1.

9.8 Linear Dependence

The definition of the linear dependence or independence of a set of vectors is essentially identical to Definition 3.2.1 for a set of functions, with the word "functions" changed to "vectors:"

DEFINITION 9.8.1 *Linear Dependence and Linear Independence*
A set of vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is said to be **linearly dependent** if at least one of them can be expressed as a linear combination of the others. If none can be so expressed, then the set is **linearly independent**.

Thus, we urge you to review Section 3.2 in conjunction with your study of this section. As in Chapter 3, we frequently use the abbreviations **LD** and **LI** to stand for linearly dependent and linearly independent, respectively.

EXAMPLE 1. Let $\mathbf{u}_1 = (1, 0)$, $\mathbf{u}_2 = (1, 1)$, and $\mathbf{u}_3 = (5, 4)$. These are LD since, by inspection, we can express $\mathbf{u}_3$ as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$: $\mathbf{u}_3 = \mathbf{u}_1 + 4\mathbf{u}_2$. (Alternatively, we could express $\mathbf{u}_2 = \frac{1}{4}\mathbf{u}_3 - \frac{1}{4}\mathbf{u}_1$, or $\mathbf{u}_1 = -4\mathbf{u}_2 + \mathbf{u}_3$). ▮

EXAMPLE 2. Let $\mathbf{u}_1 = (1, 0)$ and $\mathbf{u}_2 = (1, 1)$. These are LI since $\mathbf{u}_1$ cannot be expressed as a "linear combination of the others," namely, as a scalar multiple of $\mathbf{u}_2$, nor can $\mathbf{u}_2$ be expressed as a scalar multiple of $\mathbf{u}_1$. ▮

EXAMPLE 3. Let $\mathbf{u}_1 = (2, -1)$, $\mathbf{u}_2 = (0, 0)$, and $\mathbf{u}_3 = (0, 1)$. These are LD since we can express $\mathbf{u}_2 = 0\mathbf{u}_1 + 0\mathbf{u}_3$. (The fact that we *cannot* express $\mathbf{u}_1$ as a linear combination of $\mathbf{u}_2$ and $\mathbf{u}_3$, nor $\mathbf{u}_3$ as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$ does not alter our conclusion, for recall the words "at least one" in the definition.) ▮

It is implicit in Definition 9.8.1 that $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are all members of the same vector space; in Examples 1 to 3 that space was $\mathbb{R}^2$. Thus, it would make no sense to ask whether $\mathbf{u}_1 = (2, 5)$ and $\mathbf{u}_2 = (4, 3, 0, 1)$ are linearly dependent or not since $\mathbf{u}_1$ is a member of $\mathbb{R}^2$ while $\mathbf{u}_2$ is a member of $\mathbb{R}^4$.

The preceding examples are simple enough to be worked by inspection. In more complicated cases, the following theorem provides a systematic approach for determining whether a given vector set is linearly dependent or linearly independent.

THEOREM 9.8.1 *Test for Linear Dependence / Independence*
A finite set of vectors $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ is LD if and only if there exist scalars α_j, not all zero, such that

$$\alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0}; \tag{1}$$

if (1) holds only if all the α_j's are zero, then the set is LI.

Proof is essentially the same as for Theorem 3.2.1.

EXAMPLE 4. Consider the 4-tuples

$$\mathbf{u}_1 = (2, 0, 1, -3), \quad \mathbf{u}_2 = (0, 1, 1, 1), \quad \mathbf{u}_3 = (2, 2, 3, 0). \tag{2}$$

To see if these vectors are LI or LD, appeal directly to (1):

$$\alpha_1(2, 0, 1, -3) + \alpha_2(0, 1, 1, 1) + \alpha_3(2, 2, 3, 0) = (0, 0, 0, 0), \tag{3}$$

or $(2\alpha_1 + 2\alpha_3,\ \alpha_2 + 2\alpha_3,\ \alpha_1 + \alpha_2 + 3\alpha_3,\ -3\alpha_1 + \alpha_2) = (0, 0, 0, 0)$. Thus,

$$\begin{aligned} 2\alpha_1 \qquad\quad + 2\alpha_3 &= 0, \\ \alpha_2 + 2\alpha_3 &= 0, \\ \alpha_1 + \alpha_2 + 3\alpha_3 &= 0, \\ -3\alpha_1 + \alpha_2 \qquad\quad &= 0. \end{aligned} \tag{4}$$

Applying Gauss elimination yields

$$\begin{aligned} 2\alpha_1 \qquad + 2\alpha_3 &= 0, \\ \alpha_2 + 2\alpha_3 &= 0, \\ \alpha_3 &= 0, \\ 0 &= 0. \end{aligned} \tag{5}$$

This system admits only the trivial solution, $\alpha_1 = \alpha_2 = \alpha_3 = 0$ so $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are LI. ∎

EXAMPLE 5. Consider the 3-tuples

$$\mathbf{u}_1 = (1, 0, 1), \quad \mathbf{u}_2 = (1, 1, 1), \quad \mathbf{u}_3 = (1, 1, 2), \quad \mathbf{u}_4 = (1, 2, 1). \tag{6}$$

Working from (1), as in Example 4, we have

$$\begin{aligned} \alpha_1 + \alpha_2 + \ \alpha_3 + \ \alpha_4 &= 0, \\ \alpha_2 + \ \alpha_3 + 2\alpha_4 &= 0, \\ \alpha_1 + \alpha_2 + 2\alpha_3 + \ \alpha_4 &= 0, \end{aligned} \tag{7}$$

or, after Gauss elimination,

$$\begin{aligned} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 &= 0, \\ \alpha_2 + \alpha_3 + 2\alpha_4 &= 0, \\ \alpha_3 &= 0. \end{aligned} \tag{8}$$

This time, there exist nontrivial solutions for the α_j's so the vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are LD. [Specifically, (8) gives $\alpha_3 = 0$, $\alpha_4 = \alpha$, $\alpha_2 = -2\alpha$, $\alpha_1 = \alpha$ where α is arbitrary. With $\alpha = 1$, say, (1) becomes $\mathbf{u}_1 - 2\mathbf{u}_2 + 0\mathbf{u}_3 + \mathbf{u}_4 = \mathbf{0}$.] ▮

We conclude this section with four modest theorems, the first three being essentially the same as Theorems 3.2.4–3.2.6 for functions.

THEOREM 9.8.2 *Linear Dependence / Independence of Two Vectors*
A set of two vectors $\{\mathbf{u}_1, \mathbf{u}_2\}$ is LD if and only if one is expressible as a scalar multiple of the other.

THEOREM 9.8.3 *Linear Dependence of Sets Containing the Zero Vector*
A set containing the zero vector is LD.

THEOREM 9.8.4 *Equating Coefficients*
Let $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ be LI. Then, for

$$a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k = b_1\mathbf{u}_1 + \cdots + b_k\mathbf{u}_k$$

to hold, it is necessary and sufficient that $a_j = b_j$ for each $j = 1, \dots, k$. That is, the coefficients of corresponding vectors on the left- and right-hand sides must match.

THEOREM 9.8.5 *Orthogonal Sets*
Every finite orthogonal set of (nonzero) vectors is LI.

Proof of Theorem 9.8.5: Dot $\mathbf{u}_1$ into both sides of

$$\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k = \mathbf{0}. \tag{9}$$

In other words,

$$\begin{aligned} \mathbf{u}_1 \cdot (\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_k) &= \mathbf{u}_1 \cdot \mathbf{0}, \\ \alpha_1\mathbf{u}_1 \cdot \mathbf{u}_1 + \alpha_2\mathbf{u}_1 \cdot \mathbf{u}_2 + \cdots + \alpha_k\mathbf{u}_1 \cdot \mathbf{u}_k &= 0, \\ \alpha_1 \|\mathbf{u}_1\|^2 + 0 + \cdots + 0 &= 0. \end{aligned} \tag{10}$$

Now $\mathbf{u}_1 \neq \mathbf{0}$ implies that $\|\mathbf{u}_1\| \neq 0$ so it follows from (10) that $\alpha_1 = 0$. Similarly, dotting $\mathbf{u}_2$ into (9) gives $\alpha_2 = 0$, and so on. Since $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$, the $\mathbf{u}_j$'s must be LI, as claimed. ■

EXAMPLE 6. The set $\{(2,1),(1,5)\}$ in $\mathbb{R}^2$ is LI because neither vector can be expressed as a scalar multiple of the other. ▮

EXAMPLE 7. Let

$$\mathbf{u}_1 = (4,-1,1,2), \quad \mathbf{u}_2 = (3,0,2,5), \quad \mathbf{u}_3 = (0,0,0,0)$$

in $\mathbb{R}^4$. The set is LD, according to Theorem 9.8.3 because it contains the zero vector $\mathbf{u}_3 = \mathbf{0}$. That is, $\mathbf{u}_3$ can be expressed as a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$: $\mathbf{u}_3 = 0\mathbf{u}_1 + 0\mathbf{u}_2$. If the preceding sentence is not clear, rewrite the equation as $0\mathbf{u}_1 + 0\mathbf{u}_2 - 1\mathbf{u}_3 = \mathbf{0}$ and observe that the α_j coefficients (0, 0, and -1) are not all zero. ▮

Closure. The foregoing discussion of the linear dependence / independence of vectors is essentially the same as the discussion of the linear dependence / independence of functions in Section 3.2, except that the Wronskian determinant test did not carry over.

EXERCISES 9.8

1. (a) Can a set be neither LD nor LI? Explain.
(b) Can a set be both LD and LI? Explain.

2. Show that the following sets are LD by expressing one of the vectors as a linear combination of the others.

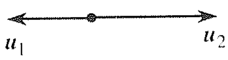
(a) $\{(1,1),(1,2),(3,4)\}$
(b) $\{(1,4),(2,8),(3,-1)\}$
(c) $\{(1,-1),(4,2),(-3,3)\}$
(d) $\{(1,2,3),(3,2,1),(5,5,5)\}$
(e) $\{(1,0,0),(0,1,0),(3,3,0),(2,-7,9)\}$

3. Determine whether the following set is LI or LD. If it is LD, then give a linear relation among the vectors.

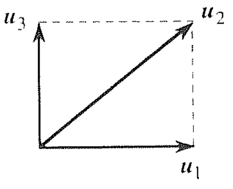
(a) $(1,3),(2,0),(-1,3),(7,3)$
(b) $(1,3),(2,0),(1,2),(-1,5)$
(c) $(2,3,0),(1,-2,3)$
(d) $(2,3,0),(1,-2,4),(1,1,0),(1,1,1)$
(e) $(0,0,2),(0,0,3),(2,-1,5),(1,2,4),(7,9,1),(2,0,-4)$
(f) $(2,3,0,0),(1,-5,0,2),(3,1,2,2)$
(g) $(1,3,2,0),(4,1,-2,-2),(0,2,0,3),(4,7,1,2)$
(h) $(2,0,1,-1,0),(1,2,0,3,1),(4,-4,3,-9,-2)$
(i) $(1,3,0),(0,1,-1),(0,0,0)$
(j) $(1,1,0,0),(1,-1,0,0),(0,0,-2,2),(0,0,1,1)$
(k) $(1,-3,0,2,1),(-2,6,0,-4,-2)$
(l) $(5,4,1,1),(0,0,0,0),(1,9,-7,2)$
(m) $(1,2,3,4),(2,3,4,5)$
(n) $(2,1,-1),(1,4,2),(3,-2,-4)$
(o) $(7,1,0),(-1,1,4),(2,3,5)$
(p) $(1,2,-1),(1,0,1),(3,-2,5)$
(q) $(3,1,0,0),(1,-2,4,1),(2,1,6,5)$
(r) $(2,4,0,1),(1,0,1,2),(1,-3,1,2),(1,1,-1-1)$

4. Show, by graphical means, that the vector sets shown below, and lying in the plane of the paper, are LD. (The emphasis here is on the method and ideas, not on graphical precision.)

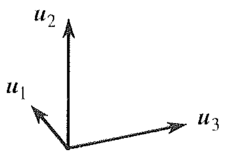
(*a*)



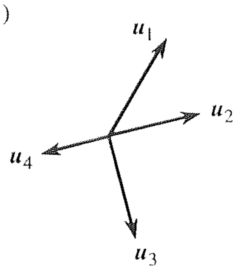
(*b*)



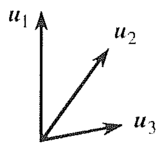
(c)



(d)



(e)



5. If $\mathbf{u}_1$ and $\mathbf{u}_2$ are LI, $\mathbf{u}_1$ and $\mathbf{u}_3$ are LI and $\mathbf{u}_2$ and $\mathbf{u}_3$ are LI, does it follow that $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ is LI? Prove or disprove.

6. Prove or disprove:

(a) $\mathbf{v}$ is in span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ if $\{\mathbf{v}, \mathbf{u}_1, \dots, \mathbf{u}_k\}$ is LD.
(b) $\mathbf{v}$ is not in span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ if $\{\mathbf{v}, \mathbf{u}_1, \dots, \mathbf{u}_k\}$ is LI.
(c) $\mathbf{v}$ is not in span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ if and only if $\{\mathbf{v}, \mathbf{u}_1, \dots, \mathbf{u}_k\}$ is LI.

7. (a) Prove Theorem 9.8.2.
(b) Prove Theorem 9.8.3.
(c) Prove Theorem 9.8.4.

9.9 Bases, Expansions, Dimension

9.9.1. Bases and expansions. In the calculus we learn that a given function $f(x)$ can be "expanded" as a linear combination of powers of x (namely $1, x, x^2, \dots$),

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots. \tag{1}$$

We call $a_0, a_1, a_2, \dots$ the "expansion coefficients," and these can be computed from $f(x)$ as $a_j = f^{(j)}(0)/j!$. Such representation of a given function is important, and examples such as $e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots$ and $\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \cdots$ are familiar to us.

Likewise useful, in Chapters 9–12, are the expansion of a given vector $\mathbf{u}$ in terms of a set of "base vectors" $\mathbf{e}_1, \dots, \mathbf{e}_k$:

$$\mathbf{u} = \alpha_1 \mathbf{e}_1 + \cdots + \alpha_k \mathbf{e}_k. \tag{2}$$

How do we come up with such sets of base vectors and, once we know the $\mathbf{e}_j$'s and the given $\mathbf{u}$, how do we compute the expansion coefficients α_j? The story is simpler than for the power series of functions because whereas (1) is an infinite series and one needs to deal with the sophisticated issue of convergence, our vector expansions in Chapters 9–12 entail only a *finite* number of terms.

Beginning simply, consider the vector space $\mathbb{R}^2$, the set of all vectors in the plane of the paper. In particular, consider the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ shown in Fig. 1a. It should be evident (Theorem 9.8.2) that $\mathbf{e}_1$ and $\mathbf{e}_2$ are LI and that they span the space so that any given vector, such as $\mathbf{u}$ in Fig. 1b and $\mathbf{v}$ in Fig. 1c, can be expressed as a linear combination of them.

For the vector $\mathbf{u}$, for example, $\mathbf{u} = \mathbf{OA} + \mathbf{OB}$; with the aid of a scale, $\mathbf{OA} = 1.6\mathbf{e}_1$ and $\mathbf{OB} = 2\mathbf{e}_2$, so that

$$\mathbf{u} = 1.6\mathbf{e}_1 + 2\mathbf{e}_2. \tag{3}$$

Similarly (Fig. 1c),

$$\mathbf{v} = 2\mathbf{e}_1 - 2.5\mathbf{e}_2, \tag{4}$$

and so on, for any given vector in the plane. Of course, the zero vector is simply $\mathbf{0} = 0\mathbf{e}_1 + 0\mathbf{e}_2$.

The formulas (3) and (4) are examples of the *expansion* of a given vector [$\mathbf{u}$ in (3), $\mathbf{v}$ in (4)] in terms of a set of *base vectors* [the set $\{\mathbf{e}_1, \mathbf{e}_2\}$].

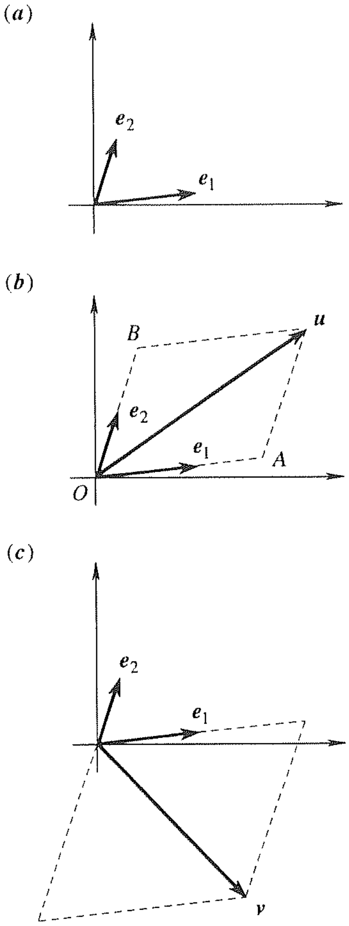


Figure 1. Vector expansion in $\mathbb{R}^2$.

DEFINITION 9.9.1 *Basis*
A finite set of vectors $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ in a vector space S is a **basis** for S if each vector $\mathbf{u}$ in S can be expressed (i.e., "expanded") uniquely in the form

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k = \sum_{j=1}^{k} \alpha_j\mathbf{e}_j. \tag{5}$$

By the expansion (5) being unique, we mean that the α_j *expansion coefficients* are uniquely determined.

THEOREM 9.9.1 *Test for Basis*
A finite set $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ in a vector space S is a basis for S if and only if it spans S and is LI.

Proof: First, it follows from the definition of the verb span that every vector $\mathbf{u}$ in S can be expanded as in (5) if and only if the set $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ spans S. Turning to the question of the uniqueness of the expansion, suppose that both expansions

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k, \tag{6}$$
$$\mathbf{u} = \beta_1\mathbf{e}_1 + \cdots + \beta_k\mathbf{e}_k \tag{7}$$

hold for any given vector $\mathbf{u}$ in S. Subtracting (7) from (6) gives

$$(\alpha_1 - \beta_1)\mathbf{e}_1 + \cdots + (\alpha_k - \beta_k)\mathbf{e}_k = \mathbf{0}. \tag{8}$$

Now, each of the coefficients $(\alpha_1 - \beta_1), \dots, (\alpha_k - \beta_k)$ in (8) must be zero, in which case $\alpha_1 = \beta_1, \dots, \alpha_k = \beta_k$ and expansions (6) and (7) are identical if and only if the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is LI. Hence, the expansion (5) is unique if and only if the set is LI, and this completes the proof. ∎

The key idea revealed in the foregoing proof is that a basis needs to contain *enough* vectors but *not too many*: enough so that the set spans the space and can therefore be used to expand any given vector in the space, but not too many, in order that such expansions will be unique.

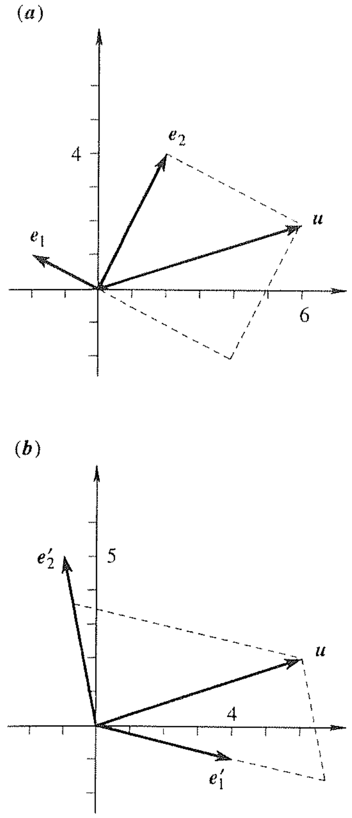


Figure 2. Two bases for $\mathbb{R}^2$.

EXAMPLE 1. Consider the vectors

$$\mathbf{e}_1 = (-2, 1), \quad \mathbf{e}_2 = (2, 4). \tag{9}$$

As may be verified, the set (9) is LI and spans $\mathbb{R}^2$ and is therefore a basis for $\mathbb{R}^2$.

Using that set to expand the vector $\mathbf{u} = (6, 2)$, say, we express

$$\mathbf{u} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2, \tag{10}$$

or $(6, 2) = (-2\alpha_1, \alpha_1) + (2\alpha_2, 4\alpha_2)$. Hence,

$$\begin{aligned} -2\alpha_1 + 2\alpha_2 &= 6, \\ \alpha_1 + 4\alpha_2 &= 2. \end{aligned} \tag{11}$$

Solving (11), $\alpha_1 = -2$ and $\alpha_2 = 1$ so the expansion (10) is

$$\mathbf{u} = -2\mathbf{e}_1 + \mathbf{e}_2, \tag{12}$$

as displayed in Fig. 2a.

It is to be emphasized that the basis (9) shown in Fig. 2a is by no means the *only* basis for $\mathbb{R}^2$; there are slews of them. For example, it is readily verified that another is

$$\mathbf{e}_1' = (4, -1), \quad \mathbf{e}_2' = (-1, 5), \tag{13}$$

and in this case the expansion of $\mathbf{u} = (6, 2)$ is found to be

$$\mathbf{u} = \frac{32}{19}\mathbf{e}_1' + \frac{14}{19}\mathbf{e}_2', \tag{14}$$

as depicted in Fig. 2b.

COMMENT. The difference between the expansions (12) and (13) is *not* at odds with the notion of uniqueness since the two expansions are with respect to different bases. In other words, (12) *is* the unique expansion of $\mathbf{u}$ in terms of the $\mathbf{e}_1, \mathbf{e}_2$ basis, and (14) *is* the unique expansion of $\mathbf{u}$ in terms of the $\mathbf{e}_1', \mathbf{e}_2'$ basis. ∎

9.9.2. Dimension. If we always worked in 2-space or 3-space, the concept of dimension would hardly need elaboration; for example, 3-space is three-dimensional, a plane within it is two-dimensional, and a line within it is one-dimensional. However, having generalized our vector concept beyond 3-space, we need to clarify the idea of dimension.

DEFINITION 9.9.2 *Dimension*
If the greatest number of LI vectors that can be found in a vector space $\mathcal{S}$ is k,

where $1 \leq k < \infty$, then $\mathcal{S}$ is **k-dimensional**, and we write

$$\dim \mathcal{S} = k.$$

If $\mathcal{S}$ is the zero vector space (i.e., if it contains only the zero vector), we define $\dim \mathcal{S} = 0$. If an arbitrarily large number of LI vectors can be found in $\mathcal{S}$, we say that $\mathcal{S}$ is **infinite-dimensional**.*

To determine the dimension of a given vector space, it may be more convenient to use the following theorem than to work directly from Definition 9.9.2.

THEOREM 9.9.2 *Test for Dimension*
If a vector space $\mathcal{S}$ admits a basis consisting of k vectors, then $\mathcal{S}$ is k-dimensional.

Proof: Let $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ be a basis for $\mathcal{S}$. Because these vectors form a basis, they must be LI. Hence, we have *at least* k LI vectors in $\mathcal{S}$, and it remains to show that *no more than* k LI vectors can be found in $\mathcal{S}$. Suppose that vectors $\mathbf{e}_1', \dots, \mathbf{e}_{k+1}'$ in $\mathcal{S}$ are LI. Each of these can be expanded in terms of the given base vectors, as

$$\begin{aligned} \mathbf{e}_1' &= a_{11}\mathbf{e}_1 + \cdots + a_{1k}\mathbf{e}_k, \\ &\vdots \\ \mathbf{e}_{k+1}' &= a_{k+1,1}\mathbf{e}_1 + \cdots + a_{k+1,k}\mathbf{e}_k, \end{aligned} \tag{15}$$

say. Putting these expressions into the equation

$$\alpha_1\mathbf{e}_1' + \alpha_2\mathbf{e}_2' + \cdots + \alpha_{k+1}\mathbf{e}_k' = \mathbf{0} \tag{16}$$

and grouping terms gives

$$(\alpha_1 a_{11} + \cdots + \alpha_{k+1}a_{k+1,1})\,\mathbf{e}_1 + \cdots + (\alpha_1 a_{1k} + \cdots + \alpha_{k+1}a_{k+1,k})\,\mathbf{e}_k = \mathbf{0}.$$

But the set $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ is LI since it is a basis, so each coefficient in the preceding equation must be zero:

$$\begin{aligned} a_{11}\alpha_1 + \cdots + a_{k+1,1}\alpha_{k+1} &= 0, \\ &\vdots \\ a_{1k}\alpha_1 + \cdots + a_{k+1,k}\alpha_{k+1} &= 0. \end{aligned} \tag{17}$$

These are k linear homogeneous equations in the $k+1$ unknowns α_1 through α_{k+1}, and such a system necessarily admits nontrivial solutions (Theorem 8.3.4). Thus, the α's in (17) are not all necessarily zero so the vectors $\mathbf{e}_1', \dots, \mathbf{e}_{k+1}'$ could

*Infinite-dimensional function spaces will be studied in Chapter 17.

not have been LI after all. Hence, it is not possible to find more than k LI vectors in S, and this completes the proof. ■

The spaces of chief concern in Chapters 9–12 are the n-tuple spaces $\mathbb{R}^n$ and subspaces thereof. For $\mathbb{R}^n$ we can say the following.

THEOREM 9.9.3 *Dimension of* $\mathbb{R}^n$
The dimension of $\mathbb{R}^n$ is n: dim $\mathbb{R}^n = n$.

Proof: The vectors

$$\begin{aligned} \mathbf{e}_1 &= (1, 0, \dots, 0), \\ \mathbf{e}_2 &= (0, 1, 0, \dots, 0), \\ &\vdots \\ \mathbf{e}_n &= (0, \dots, 0, 1) \end{aligned} \tag{18}$$

constitute a basis for $\mathbb{R}^n$ because any vector $\mathbf{u} = (u_1, \dots, u_n)$ in $\mathbb{R}^n$ can be expanded uniquely as $\mathbf{u} = u_1\mathbf{e}_1 + \cdots + u_n\mathbf{e}_n$. Since this basis contains n vectors, it follows from Theorem 9.9.2 that $\mathbb{R}^n$ is n-dimensional. ■

Indeed, we might well have questioned the reasonableness of our definition of dimension if $\mathbb{R}^n$ had turned out to be other than n-dimensional! The ON basis (18) is called the **standard basis** for $\mathbb{R}^n$ (and is the n-space generalization of the "$\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$" ON basis that might be known to you from other courses).

Finally, what about the dimension of a *subspace*, for example, the subspace of $\mathbb{R}^3$ that is spanned by two given vectors?

THEOREM 9.9.4 *Dimension of Span* $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$
The dimension of span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$, where the $\mathbf{u}_j$'s are not all zero, denoted as dim [span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$], is equal to the greatest number of LI vectors within the generating set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$.

Proof: Denote the generating set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ as U. Let the greatest number of LI vectors in U be N, where $1 \leq N \leq k$. It may be assumed, without loss of generality, that the members of U have been numbered so that $\mathbf{u}_1, \dots, \mathbf{u}_N$ are LI. Then each of the remaining members of U, namely $\mathbf{u}_{N+1}, \dots, \mathbf{u}_k$, can be expressed as a linear combination of $\mathbf{u}_1, \dots, \mathbf{u}_N$. Surely, then, each vector in span U can similarly be expressed as a linear combination of $\mathbf{u}_1, \dots, \mathbf{u}_N$. Now $\{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ is LI and spans span U. According to Theorem 9.9.3, then, the dimension of span U is N; that is, it is the same as the greatest number of LI vectors in U, as was to be proved. ■

EXAMPLE 2. Let

$$\mathbf{u}_1 = (3, -1, 2, 1), \quad \mathbf{u}_2 = (1, 1, 0, -1), \quad \mathbf{u}_3 = (4, 0, 2, 0).$$

These vectors are, of course, members of $\mathbb{R}^4$. But since $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ are only three vectors, dim [span $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$] is *at most* three. In fact, it is *not* three since we see that $\mathbf{u}_3 = \mathbf{u}_1 + \mathbf{u}_2$. But $\mathbf{u}_1$ and $\mathbf{u}_2$, say, are LI since neither is a scalar multiple of the other. Thus, there are only two LI vectors within the generating set so dim [span $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$] = 2. ▮

In Example 2 we determined that the greatest number of LI vectors in the generating set was 2 by inspection. What if we wish to determine dim [span $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$] where the $\mathbf{u}_j$'s are members of $\mathbb{R}^8$, and $k = 6$, say? For such a large problem we cannot expect "inspection" to work. Yet, what are we to do, test the $\mathbf{u}_j$'s for linear independence one at a time, two at a time, three at a time, and so on, until we determine the greatest number of LI vectors in $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$? That would be quite tedious. No, we will see later, in Chapter 10, that the best way to determine the greatest number of LI vectors in a given set is to determine the "rank" of a certain matrix, and that can be done by the extremely efficient method of elementary row operations. Meanwhile, in the present section, we "get by" by keeping the examples and exercises simple enough so that we can rely on inspection.

Let us return, now, to our discussion of bases and expansions.

9.9.3. Orthogonal bases. If, as in Example 1, there are many bases for a given space, then how do we decide which one to select? We will find that in most applications the most convenient basis to use is dictated by the context, so let us not worry about that now. This point is addressed in Chapter 11 as well as in the chapters on PDEs.

However, we do wish to show, here, that *orthogonal* bases are to be preferred whenever possible. For observe from Example 1 that to expand $\mathbf{u}$ (that is, to compute the α_j expansion coefficients) we needed to solve the system (11) of two equations in two unknowns. Similarly, if we seek to expand a given vector in $\mathbb{R}^8$, then there will be eight base vectors (because $\mathbb{R}^8$ is eight-dimensional) and eight α_j expansion coefficients, and these will be found by solving a system [analogous to (11)] of eight equations in the eight unknown α_j's. Thus, the expansion process can be quite laborious.

On the other hand, suppose that $\{\mathbf{e}_1, \dots, \mathbf{e}_k\}$ is an *orthogonal* basis for $\mathcal{S}$; that is, it is not only a basis but also happens to be an orthogonal set:

$$\mathbf{e}_i \cdot \mathbf{e}_j = 0 \quad \text{if} \quad i \neq j. \tag{19}$$

Suppose that we wish to expand a given vector $\mathbf{u}$ in $\mathcal{S}$ in terms of that basis; that is, we wish to determine the coefficients $\alpha_1, \dots, \alpha_k$ in the expansion

$$\mathbf{u} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \cdots + \alpha_k \mathbf{e}_k. \tag{20}$$

To accomplish this, dot (20) with $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_k$, in turn. Doing so, and using

(19), we obtain the linear system

$$\begin{aligned}
\mathbf{u}\cdot\mathbf{e}_1 &= (\mathbf{e}_1\cdot\mathbf{e}_1)\alpha_1 + 0\alpha_2 + \cdots + 0\alpha_k,\\
\mathbf{u}\cdot\mathbf{e}_2 &= 0\alpha_1 + (\mathbf{e}_2\cdot\mathbf{e}_2)\alpha_2 + 0\alpha_3 + \cdots + 0\alpha_k,\\
&\ \ \vdots\\
\mathbf{u}\cdot\mathbf{e}_k &= 0\alpha_1 + \cdots + 0\alpha_{k-1} + (\mathbf{e}_k\cdot\mathbf{e}_k)\alpha_k,
\end{aligned} \tag{21}$$

where all of the quantities $\mathbf{u}\cdot\mathbf{e}_1,\ldots,\mathbf{u}\cdot\mathbf{e}_k, \mathbf{e}_1\cdot\mathbf{e}_1,\ldots,\mathbf{e}_k\cdot\mathbf{e}_k$ are computable since $\mathbf{u},\mathbf{e}_1,\ldots,\mathbf{e}_k$ are known. The crucial point is that even though (21) is still k equations in the k unknown α_j's, the system is *uncoupled* (i.e., the only unknown in the first equation is α_1, the only one in the second is α_2, and so on) and readily gives

$$\alpha_1 = \frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}, \quad \alpha_2 = \frac{\mathbf{u}\cdot\mathbf{e}_2}{\mathbf{e}_2\cdot\mathbf{e}_2}, \quad \ldots, \quad \alpha_k = \frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}, \tag{22}$$

provided, of course, that none of the denominators vanish. But these quantities cannot vanish because $\mathbf{e}_j\cdot\mathbf{e}_j = \|\mathbf{e}_j\|^2$, which is zero if and only if $\mathbf{e}_j = \mathbf{0}$, and this cannot be because if any $\mathbf{e}_j$ were $\mathbf{0}$, then the set $\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}$ would be LD (Theorem 9.8.3), and hence not a basis.

Thus, *if the* $\{\mathbf{e}_1,\ldots,\mathbf{e}_k\}$ *basis is orthogonal, the expansion of any given* $\mathbf{u}$ *is simply*

$$\boxed{\mathbf{u} = \left(\frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}\right)\mathbf{e}_1 + \cdots + \left(\frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}\right)\mathbf{e}_k = \sum_{j=1}^{k}\left(\frac{\mathbf{u}\cdot\mathbf{e}_j}{\mathbf{e}_j\cdot\mathbf{e}_j}\right)\mathbf{e}_j.} \tag{23}$$

If, besides being orthogonal, the $\mathbf{e}_j$'s are normalized ($\|\mathbf{e}_j\| = 1$) so that they constitute an ON (orthonormal) basis, then (23) simplifies slightly to

$$\boxed{\mathbf{u} = (\mathbf{u}\cdot\hat{\mathbf{e}}_1)\,\hat{\mathbf{e}}_1 + \cdots + (\mathbf{u}\cdot\hat{\mathbf{e}}_k)\,\hat{\mathbf{e}}_k = \sum_{j=1}^{k}(\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j,} \tag{24}$$

where we recall that carets denote unit vectors.

EXAMPLE 3. Expand $\mathbf{u} = (4,3,-3,6)$ in terms of the orthogonal base vectors $\mathbf{e}_1 = (1,0,2,0)$, $\mathbf{e}_2 = (0,1,0,0)$, $\mathbf{e}_3 = (-2,0,1,5)$, $\mathbf{e}_4 = (-2,0,1,-1)$ of $\mathbb{R}^4$. This basis is orthogonal but not ON so we use (23) rather than (24). Computing $\mathbf{u}\cdot\mathbf{e}_1 = -2$, $\mathbf{e}_1\cdot\mathbf{e}_1 = 5$, and so on, (23) gives

$$\mathbf{u} = -\frac{2}{5}\mathbf{e}_1 + 3\mathbf{e}_2 + \frac{19}{30}\mathbf{e}_3 - \frac{17}{6}\mathbf{e}_4. \tag{25}$$

Alternatively, we could have inferred, from $\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_4\mathbf{e}_4$, the four equations

$$\begin{aligned} \alpha_1 \qquad - 2\alpha_3 - 2\alpha_4 &= 4, \\ \alpha_2 \qquad\qquad\qquad &= 3, \\ 2\alpha_1 \qquad + \alpha_3 + \alpha_4 &= -3, \\ 5\alpha_3 - \alpha_4 &= 6 \end{aligned} \tag{26}$$

on the four unknown α_j's, and solved these by Gauss elimination, but it is much easier to "cash in" on the orthogonality of the basis and to use (23). If we choose to work with an ON basis, we can scale the $\mathbf{e}_j$'s as $\hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}(1,0,2,0)$, $\hat{\mathbf{e}}_2 = (0,1,0,0)$, $\hat{\mathbf{e}}_3 = \frac{1}{\sqrt{30}}(-2,0,1,5)$, $\hat{\mathbf{e}}_4 = \frac{1}{\sqrt{6}}(-2,0,1,-1)$. Then (24) gives

$$\mathbf{u} = -\frac{2}{\sqrt{5}}\hat{\mathbf{e}}_1 + 3\hat{\mathbf{e}}_2 + \frac{19}{\sqrt{30}}\hat{\mathbf{e}}_3 - \frac{17}{\sqrt{6}}\hat{\mathbf{e}}_4, \tag{27}$$

which result is equivalent to (25). ▮

Given a *non*orthogonal basis there are three possibilities. First, one can use it and face up to the tedious expansion process. Second, one can "trade the nonorthogonal basis in" for an orthogonal basis using the **Gram–Schmidt orthogonalization procedure**, which procedure is introduced briefly in the exercises and discussed in detail in the next section. Third, one can retain the nonorthogonal basis but streamline the expansion process by computing and utilizing a set of **dual**, or **reciprocal**, **vectors** corresponding to the given basis, as described in the exercises.

Closure. This section is about the **expansion** of vectors, in a given vector space $\mathcal{S}$, in terms of a set of *base vectors*. A set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ in $\mathcal{S}$ is a basis for $\mathcal{S}$ if each vector $\mathbf{u}$ in $\mathcal{S}$ *can* be expanded as a *unique* linear combination of the $\mathbf{e}_j$'s. We showed (Theorem 9.9.1) that $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is indeed a basis for $\mathcal{S}$ if and only if it spans $\mathcal{S}$ (so each $\mathbf{u}$ *can* be expanded) and is LI (so the expansion is *unique*). The number of vectors in any basis for $\mathcal{S}$ is called the *dimension* of $\mathcal{S}$. For instance, $\mathbb{R}^n$ admits the *standard basis* (18), comprised of n vectors, so $\mathbb{R}^n$ is n-dimensional. And the greatest number of LI vectors in a set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ is the dimension of their span.

We found that the expansion process (i.e., the determination of the expansion coefficients) can be quite laborious if there are many base vectors, but is extremely simple if the basis is orthogonal, or ON, in which case the expansions are given by (23) or (24), respectively. You should remember those two formulas and be able to derive them as well.

EXERCISES 9.9

1. Show whether the following is a basis.

(a) $(1,0),(1,1),(1,2)$ for $\mathbb{R}^2$
(b) $(3,2),(-1,-5)$ for $\mathbb{R}^2$
(c) $(1,1)$ for $\mathbb{R}^2$
(d) $(2,0,1),(5,-1,2),(1,-1,0)$ for $\mathbb{R}^3$
(e) $(5,-1,2),(2,0,1),(1,-1,1)$ for $\mathbb{R}^3$
(f) $(2,1,0,6),(7,-1,-2,3),(4,3,2,1)$ for $\mathbb{R}^4$
(g) $(4,3,-2,1),(5,0,0,0),(2,1,-3,0),(1,2,4,5)$ for $\mathbb{R}^4$
(h) $(4,2,0,0),(1,2,3,0),(5,-2,3,1),(0,-6,0,1)$ for $\mathbb{R}^4$
(i) $(1,0,0,0),(1,1,0,0),(1,1,1,0),(1,1,1,1)$ for $\mathbb{R}^4$
(j) $(3,0,0,1),(2,0,0,1),(1,3,5,6),(4,-2,1,3)$ for $\mathbb{R}^4$
(k) $(1,3,-1,2),(1,2,4,3),(2,5,3,5),(3,7,7,8)$ for $\mathbb{R}^4$
(l) $(2,3,5,0),(1,-1,2,3),(4,1,2,3),(5,4,1,0),(1,2,4,6)$ for $\mathbb{R}^4$
(m) $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1),(0,0,0,0)$ for $\mathbb{R}^4$
(n) $(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,0)$ for $\mathbb{R}^4$
(o) $(1,1,2),(4,-2,-1)$ for span $\{(2,-4,-5),(1,-1,-1)\}$
(p) $(1,1,2),(4,-2,-1)$ for span $\{(3,-5,-6),(1,2,1)\}$
(q) $(1,1,1),(1,-1,2)$ for span $\{(2,4,1),(1,7,-2)\}$
(r) $(1,2,3),(1,0,4)$ for span $\{(3,2,0),(1,1,-1)\}$

2. Expand each vector $\mathbf{u}$ in terms of the orthogonal basis $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$ of $\mathbb{R}^3$, where $\mathbf{e}_1=(2,1,3)$, $\mathbf{e}_2=(1,-2,0)$, $\mathbf{e}_3=(6,3,-5)$.

(a) $\mathbf{u}=(9,-2,4)$ (b) $\mathbf{u}=(1,0,0)$
(c) $\mathbf{u}=(0,1,5)$ (d) $\mathbf{u}=(3,-1,1)$
(e) $\mathbf{u}=(0,5,0)$ (f) $\mathbf{u}=(1,2,3)$

3. (a)–(f) Expand each of the $\mathbf{u}$ vectors in Exercise 2 in terms of the ON basis $\{\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3\}$ of $\mathbb{R}^3$, where $\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3$ are normalized versions of $\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3$ given in Exercise 2.

4. Expand each vector $\mathbf{u}$ in terms of the orthogonal basis $\{\mathbf{e}_1,\dots,\mathbf{e}_4\}$ of $\mathbb{R}^4$, where $\mathbf{e}_1=(2,0,-1,-5)$, $\mathbf{e}_2=(2,0,-1,1)$, $\mathbf{e}_3=(0,1,0,0)$, $\mathbf{e}_4=(1,0,2,0)$.

(a) $\mathbf{u}=(1,0,0,0)$ (b) $\mathbf{u}=(0,6,0,0)$
(c) $\mathbf{u}=(2,5,1,-3)$ (d) $\mathbf{u}=(4,3,-2,0)$
(e) $\mathbf{u}=(1,2,0,5)$ (f) $\mathbf{u}=(2,-7,4,1)$
(g) $\mathbf{u}=(0,0,0,9)$ (h) $\mathbf{u}=(2,3,-2,1)$
(i) $\mathbf{u}=(0,0,5,0)$ (j) $\mathbf{u}=(1,1,1,1)$

5. Verify that the $\{\mathbf{e}_1,\dots,\mathbf{e}_4\}$ vectors given in Example 3 are a basis for $\mathbb{R}^4$. Also, solve (26) by Gauss elimination and verify that the α_j's thus obtained agree with those given in (25).

6. If $\{\mathbf{e}_1,\dots,\mathbf{e}_k\}$ is an orthogonal set in a vector space $\mathcal{S}$, is it a basis

(a) for $\mathcal{S}$? (b) for span $\{\mathbf{e}_1,\dots,\mathbf{e}_k\}$?

7. (*Zero vector space*) Show that a **zero vector space** (i.e., a vector space consisting of the zero vector alone) has no basis.

8. Let $\mathbf{u}_1=(1,0,0)$, $\mathbf{u}_2=(0,1,0)$, $\mathbf{u}_3=(0,0,1)$, $\mathbf{u}_4=(1,1,0)$, $\mathbf{u}_5=(0,1,1)$, $\mathbf{u}_6=(1,1,1)$, and $\mathbf{u}_7=($
Evaluate each of the following.

(a) dim [span $\{\mathbf{u}_1\}$]
(b) dim [span $\{\mathbf{u}_1,\mathbf{u}_2\}$]
(c) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\}$]
(d) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4\}$]
(e) dim [span $\{\mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_4\}$]
(f) dim [span $\{\mathbf{u}_1,\mathbf{u}_4,\mathbf{u}_5\}$]
(g) dim [span $\{\mathbf{u}_5,\mathbf{u}_6,\mathbf{u}_7\}$]
(h) dim [span $\{\mathbf{u}_1,\mathbf{u}_5,\mathbf{u}_6,\mathbf{u}_7\}$]

9. Let $\mathbf{u}_1=(1,0,0,0)$, $\mathbf{u}_2=(1,1,0,0)$, $\mathbf{u}_3=(1,1,1,0)$, $\mathbf{u}_4=(1,1,1,1)$, $\mathbf{u}_5=(0,0,0,1)$, $\mathbf{u}_6=(3,3,3,3)$. Evaluate each of the following.

(a) dim [span $\{\mathbf{u}_1,\mathbf{u}_3,\mathbf{u}_5\}$]
(b) dim [span $\{\mathbf{u}_1,\mathbf{u}_4,\mathbf{u}_6\}$]
(c) dim [span $\{\mathbf{u}_2,\mathbf{u}_4,\mathbf{u}_6\}$]
(d) dim [span $\{\mathbf{u}_5\}$]
(e) dim [span $\{\mathbf{u}_3,\mathbf{u}_4\}$]
(f) dim [span $\{\mathbf{u}_3,\mathbf{u}_4,\mathbf{u}_5,\mathbf{u}_6\}$]
(g) dim [span $\{\mathbf{u}_4,\mathbf{u}_6\}$]
(h) dim [span $\{\mathbf{u}_2,\mathbf{u}_3,\mathbf{u}_4,\mathbf{u}_5,\mathbf{u}_6\}$]

10. (a)–(f) Determine the dimension of the solution space in Exercise 4 of Section 9.7.

11. (*Gram–Schmidt orthogonalization process*) Given k LI vectors $\mathbf{v}_1,\dots,\mathbf{v}_k$, it is possible to obtain from them k ON vectors, say $\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_k$, in span $\{\mathbf{v}_1,\dots,\mathbf{v}_k\}$ by the **Gram–Schmidt process**, after *Jörgen P. Gram* (1850–1916) and *Erhardt Schmidt* (1876–1959), by taking $\mathbf{e}_1$ equal to $\mathbf{v}_1$, taking $\mathbf{e}_2$ equal to a suitable linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, taking $\mathbf{e}_3$ equal to a suitable linear combination of $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$, and so on, and then normalizing the results. The resulting ON set is as follows:

$$\begin{aligned}\hat{\mathrm{e}}_1 &= \frac{\mathrm{v}_1}{\|\mathrm{v}_1\|},\\ \hat{\mathrm{e}}_2 &= \frac{\mathrm{v}_2 - (\mathrm{v}_2 \cdot \hat{\mathrm{e}}_1)\hat{\mathrm{e}}_1}{\|\mathrm{v}_2 - (\mathrm{v}_2 \cdot \hat{\mathrm{e}}_1)\hat{\mathrm{e}}_1\|},\\ &\vdots\\ \hat{\mathrm{e}}_j &= \frac{\mathrm{v}_j - \sum_{i=1}^{j-1}(\mathrm{v}_j \cdot \hat{\mathrm{e}}_i)\hat{\mathrm{e}}_i}{\left\|\mathrm{v}_j - \sum_{i=1}^{j-1}(\mathrm{v}_j \cdot \hat{\mathrm{e}}_i)\hat{\mathrm{e}}_i\right\|} \quad \text{through } j = k.\end{aligned} \tag{11.1}$$

We now state the problem: Verify that each $\hat{\mathrm{e}}_j$ defined by (11.1) is a linear combination of $\mathrm{v}_1, \ldots, \mathrm{v}_j$, and that the $\hat{\mathrm{e}}_j$'s are ON. [In verifying that $\|\hat{\mathrm{e}}_j\| = 1$, be sure to show that each denominator in (11.1) is nonzero.]

12. In each case use the Gram–Schmidt formula (11.1) in Exercise 11 to obtain an ON set from the given LI set.

(a) $(4, 0), (2, 1)$
(b) $(1, -2), (3, 4)$
(c) $(1, 0, 0), (1, 1, 0), (1, 1, 1)$
(d) $(1, 1, 0), (2, -1, 1), (1, 0, 3)$
(e) $(1, 1, 1), (2, 0, -1)$
(f) $(1, 1, 1), (1, 0, 1), (1, 1, 0)$
(g) $(1, 2, 1), (1, -1, 2), (-1, 3, 1)$
(h) $(2, 0, 1), (1, 1, 1), (-2, 0, 3)$
(i) $(2, 1, 1, 0), (1, 5, -1, 2)$
(j) $(6, -1, 1, 2, 1), (2, 3, -1, 1, 4)$

13. (*The dual or reciprocal vectors*) For definiteness, consider our vector space $\mathcal{S}$ to be $\mathbb{R}^n$.

(a) If $\{\hat{\mathrm{e}}_1, \ldots, \hat{\mathrm{e}}_n\}$ is an ON basis for $\mathbb{R}^n$, and $\mathbf{u}$ is in $\mathbb{R}^n$, then by dotting $\hat{\mathrm{e}}_i$, into both sides of the equation $\mathbf{u} = \sum_{j=1}^{n} \alpha_j \hat{\mathrm{e}}_j$, we find that $\alpha_i = \mathbf{u} \cdot \hat{\mathrm{e}}_i$ so that the expansion of $\mathbf{u}$ in terms of the given basis is

$$\mathbf{u} = \sum_{j=1}^{n} (\mathbf{u} \cdot \hat{\mathrm{e}}_j)\hat{\mathrm{e}}_j. \tag{13.1}$$

If, instead, we have a basis $\{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$ which is *not* ON, then, as noted in the text, the expansion process is not so simple. However, suppose that we can find a set $\{\mathrm{e}_1^*, \ldots, \mathrm{e}_n^*\}$ such that

$$\mathrm{e}_i \cdot \mathrm{e}_j^* = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases} \tag{13.2}$$

Then show that dotting e_i^* into $\mathbf{u} = \sum_{j=1}^{n} \alpha_j \mathrm{e}_j$ gives $\alpha_i = \mathbf{u} \cdot \mathrm{e}_i^*$ so that

$$\mathbf{u} = \sum_{j=1}^{n} (\mathbf{u} \cdot \mathrm{e}_j^*)\mathrm{e}_j. \tag{13.3}$$

The set $\{\mathrm{e}_1^*, \ldots, \mathrm{e}_n^*\}$ is called the **dual**, or **reciprocal**, **set** corresponding to the original set $\{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$. (We will see in the last exercise in Section 10.6 that the dual set exists, is unique, and is itself a basis for $\mathbb{R}^n$, the so-called **dual** or **reciprocal basis**.)

(b) Given the basis $\mathrm{e}_1 = (1, 0), \mathrm{e}_2 = (1, 1)$ for $\mathbb{R}^2$, use equation (13.2) to determine the dual vectors $\mathrm{e}_1^*, \mathrm{e}_2^*$. Then use equation (13.3) to expand $\mathbf{u} = (3, 1)$. Sketch $\mathrm{e}_1, \mathrm{e}_2, \mathrm{e}_1^*, \mathrm{e}_2^*, \mathbf{u}$ to scale, and verify the expansion graphically, that is, by means of the parallelogram rule of vector addition.

(c) Repeat part (b), for $\mathrm{e}_1 = (2, 1), \mathrm{e}_2 = (0, 2), \mathbf{u} = (-3, 2)$.

(d) Repeat part (b), for $\mathrm{e}_1 = (-1, 1), \mathrm{e}_2 = (2, 1), \mathbf{u} = (0, 4)$.

(e) Given the basis $\mathrm{e}_1 = (1, 0, 0), \mathrm{e}_2 = (1, 1, 0), \mathrm{e}_3 = (1, 1, 1)$ for $\mathbb{R}^3$, use equation (13.2) to determine the dual vectors $\mathrm{e}_1^*, \mathrm{e}_2^*, \mathrm{e}_3^*$. Then use equation (13.3) to expand each of the vectors $\mathbf{u} = (4, -1, 5), \mathbf{v} = (0, 0, 2), \mathbf{w} = (5, -2, 3)$. Be sure to see that the dual vectors get computed once and for all, for a given basis $\{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$; once we have got them, expansions of the form (13.3) are simple.

(f) Repeat part (e) for $\mathrm{e}_1 = (2, 0, 1), \mathrm{e}_2 = (1, 1, 0), \mathrm{e}_3 = (1, -1, 3)$, and $\mathbf{u} = (6, 1, 0), \mathbf{v} = (1, 2, 4), \mathbf{w} = (0, 3, 0)$.

(g) Show that if the $\{\mathrm{e}_1, \ldots, \mathrm{e}_n\}$ basis does happen to be ON, then the dual vectors coalesce with the e_j's, i.e., $\mathrm{e}_j^* = \mathrm{e}_j$ for $j = 1, 2, \ldots, n$.

9.10 Best Approximation

Let $\mathcal{S}$ be a normed inner product vector space (i.e., a vector space with both a norm and an inner, or dot, product defined), and let the norm be the "natural norm" $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$. We know that if $\{\mathbf{e}_1,\dots,\mathbf{e}_N\}$ is a basis for $\mathcal{S}$, then any vector $\mathbf{u}$ in $\mathcal{S}$ can be (uniquely) expanded in the form $\mathbf{u} = \sum_{j=1}^{N} c_j\mathbf{e}_j$. If the basis is orthogonal, then the expansion process is easy, with the c_j's computed, from the given vector $\mathbf{u}$ and the base vectors $\mathbf{e}_j$, as $c_j = (\mathbf{u}\cdot\mathbf{e}_j)/(\mathbf{e}_j\cdot\mathbf{e}_j)$. And if the basis is not only orthogonal but ON, then $\mathbf{u} = \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j$, where $c_j = \mathbf{u}\cdot\hat{\mathbf{e}}_j$.

However, what if we do not have a "full deck?" That is, what if $\{\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_N\}$ is ON, but falls short of being a basis for $\mathcal{S}$ (i.e., $N <$ dim $\mathcal{S}$)? If $\mathbf{u}$ happens to fall within span $\{\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_N\}$, which subspace of $\mathcal{S}$ we denote as $\mathcal{T}$, then it can still be expanded in terms of $\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_N$, but if it is not in $\mathcal{T}$, then it cannot be so expanded.

In the latter case the question arises, what is the best *approximation* of $\mathbf{u}$ in terms of $\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_N$? In this section we answer that question in general, and illustrate the results for the case where $\mathcal{S}$ is $\mathbb{R}^n$. Later in this book, when we study Fourier series and partial differential equations, our interest will be in function spaces instead.

9.10.1. Best approximation and orthogonal projection. The *best approximation problem*, which we address is this: given a vector $\mathbf{u}$ in $\mathcal{S}$, and an ON set $\{\hat{\mathbf{e}}_1,\dots,\hat{\mathbf{e}}_N\}$ in $\mathcal{S}$, what is the best approximation

$$\mathbf{u} \approx c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N = \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j? \tag{1}$$

That is, how do we compute the c_j coefficients so as to render the error vector $\mathbf{E} = \mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j$ as small as possible? In other words, how do we choose the c_j's so as to minimize the norm of the error vector $\|\mathbf{E}\|$? If $\|\mathbf{E}\|$ is a minimum, then so is $\|\mathbf{E}\|^2$, so let us minimize $\|\mathbf{E}\|^2$ (to avoid square roots), where

$$\begin{aligned}\|\mathbf{E}\|^2 = \mathbf{E}\cdot\mathbf{E} &= \left(\mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j\right)\cdot\left(\mathbf{u} - \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j\right)\\ &= \mathbf{u}\cdot\mathbf{u} - 2\sum_{j=1}^{N} c_j\,(\mathbf{u}\cdot\hat{\mathbf{e}}_j) + \sum_{j=1}^{N} c_j^2,\end{aligned} \tag{2}$$

and where the step

$$\begin{aligned}\left(\sum_{1}^{N} c_j\hat{\mathbf{e}}_j\right)\cdot\left(\sum_{1}^{N} c_j\hat{\mathbf{e}}_j\right) &= (c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N)\cdot(c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N)\\ &= c_1^2 + \cdots + c_N^2 = \sum_{1}^{N} c_j^2\end{aligned} \tag{3}$$

follows from the orthonormality of the $\hat{e}_j$'s.

Defining $\mathbf{u}\cdot\hat{e}_j \equiv \alpha_j$ and noting that $\mathbf{u}\cdot\mathbf{u} = \|\mathbf{u}\|^2$, we may express (2) as

$$\|\mathbf{E}\|^2 = \sum_{j=1}^{N} c_j^2 - 2\sum_{j=1}^{N} \alpha_j c_j + \|\mathbf{u}\|^2,$$

or, completing the square, as

$$\|\mathbf{E}\|^2 = \sum_{j=1}^{N} (c_j - \alpha_j)^2 + \|\mathbf{u}\|^2 - \sum_{j=1}^{N} \alpha_j^2. \tag{4}$$

Observe that $\mathbf{u}$ and the ON set $\{\hat{e}_1, \dots, \hat{e}_N\}$ are given so that $\|\mathbf{u}\|$ and the α_j's in (4) are fixed computable quantities: $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$ and $\alpha_j = \mathbf{u}\cdot\hat{e}_j$ for $j = 1, 2, \dots, N$. Thus, in seeking to minimize the right-hand side of (4), the only control we exercise is in our choice of the c_j's. The right-hand side of (4) is greater than or equal to zero,* and so is the $\sum_{j=1}^{N} (c_j - \alpha_j)^2$ term containing the c_j's. Thus, the best that we can do is to set $c_j = \alpha_j$ ($j = 1, 2, \dots, N$). With that choice, our best approximation (1) becomes

$$\boxed{\mathbf{u} \approx \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{e}_j)\,\hat{e}_j.} \tag{5}$$

Let us summarize these results.

THEOREM 9.10.1 *Best Approximation*
Let $\mathbf{u}$ be any vector in a normed inner product vector space $\mathcal{S}$ with natural norm ($\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$), and let $\{\hat{e}_1, \dots, \hat{e}_N\}$ be an ON set in $\mathcal{S}$. Then the best approximation (1) is obtained when the c_j's are given by $c_j = \mathbf{u}\cdot\hat{e}_j$, as indicated in (5).

EXAMPLE 1. Let $\mathcal{S}$ be $\mathbb{R}^2$, $N = 1$, $\hat{e}_1 = \frac{1}{13}(12, 5)$, and $\mathbf{u} = (1, 1)$, as shown in Fig. 1. Find the best approximation $\mathbf{u} \approx c_1\hat{e}_1$, that is, the best approximation of $\mathbf{u}$ in span$\{\hat{e}_1\}$ (which is the line L). Theorem 9.10.1 gives $c_1 = \mathbf{u}\cdot\hat{e}_1 = 17/13$, and hence the best approximation

$$\mathbf{u} \approx \frac{17}{13}\hat{e}_1, \tag{6}$$

which is the vector $\mathbf{OA}$ in Fig. 1.

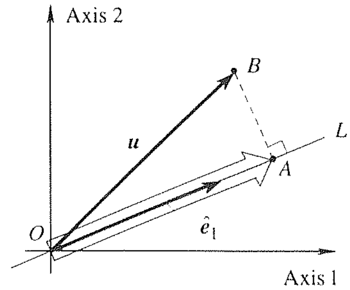


Figure 1. Best approximation of $\mathbf{u}$ in span$\{\hat{e}_1\}$.

COMMENT. Observe from the figure that the best approximation $\mathbf{OA}$ is the *orthogonal projection* of $\mathbf{u}$ onto span$\{\mathbf{e}_1\}$, which orthogonality is verified by the calculation

*This fact may not be obvious due to the minus sign in front of the last summation. But remember that the right-hand side of (4) is equal to $\|\mathbf{E}\|^2$, and surely $\|\mathbf{E}\|^2 \geq 0$.

$\mathbf{AB}\cdot\hat{\mathbf{e}}_1 = (\mathbf{u}-\mathbf{OA})\cdot\hat{\mathbf{e}}_1 = (\mathbf{u}-\frac{17}{13}\hat{\mathbf{e}}_1)\cdot\hat{\mathbf{e}}_1 = \frac{17}{13}-\frac{17}{13} = 0$. That result makes perfect sense since if $c_1\hat{\mathbf{e}}_1$ is to be the best approximation to $\mathbf{u}$, then the distance from the tip of $\mathbf{u}$ to the tip of $c_1\hat{\mathbf{e}}_1$ (which is some point on L) should be as small as possible. That shortest distance is the perpendicular distance from the tip of $\mathbf{u}$ to the line L. ▮

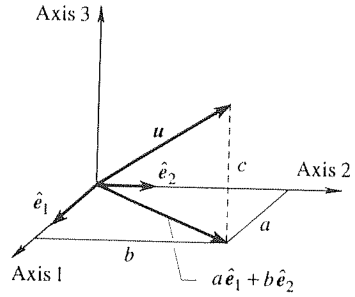


Figure 2. Best approximation of $\mathbf{u}$ in span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$.

EXAMPLE 2. Let $\mathcal{S}$ be $\mathbb{R}^3$, let $N = 2$ with $\hat{\mathbf{e}}_1 = (1,0,0)$ and $\hat{\mathbf{e}}_2 = (0,1,0)$, and let $\mathbf{u} = (a,b,c)$, as shown in Fig. 2. Computing the coefficients in (5) as $\mathbf{u}\cdot\hat{\mathbf{e}}_1 = a$ and $\mathbf{u}\cdot\hat{\mathbf{e}}_2 = b$, (5) becomes

$$\mathbf{u} \approx a\hat{\mathbf{e}}_1 + b\hat{\mathbf{e}}_2. \tag{7}$$

The latter is an equality if $c = 0$. That is, (7) is an equality if $\mathbf{u}$ happens to lie in span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$, but if $c \neq 0$ then the best approximation $a\hat{\mathbf{e}}_1 + b\hat{\mathbf{e}}_2$ to $\mathbf{u}$ is the orthogonal projection of $\mathbf{u}$ onto span $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$. ▮

In Examples 1 and 2, $\mathcal{S}$ was $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively, so we were able to draw useful pictures. In each case we discovered that the best approximation of $\mathbf{u}$ on the subspace $\mathcal{T}$ of $\mathcal{S}$ spanned by $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N$ was the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$. Is that result true in all cases? That is, is the error vector $\mathbf{E}$ necessarily orthogonal to $\mathcal{T}$? Since the error vector is

$$\mathbf{E} = \mathbf{u} - \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j, \tag{8}$$

we have

$$\begin{aligned} \mathbf{E}\cdot\hat{\mathbf{e}}_k &= \left[\mathbf{u} - \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j\right]\cdot\hat{\mathbf{e}}_k \\ &= \mathbf{u}\cdot\hat{\mathbf{e}}_k - (\mathbf{u}\cdot\hat{\mathbf{e}}_k)(1) = 0 \end{aligned} \tag{9}$$

for each $k = 1, 2, \dots, N$, where the second equality follows from the fact that $\hat{\mathbf{e}}_j\cdot\hat{\mathbf{e}}_k = 0$ if $j \neq k$ and 1 if $j = k$.

Since $\mathbf{E}$ is orthogonal to every one of the $\hat{\mathbf{e}}_j$'s, it is therefore orthogonal to every vector in $\mathcal{T}$. In that sense we say that the right-hand side of (5) is the **orthogonal projection of u onto** $\mathcal{T}$, and denote it as $\text{proj}_{\mathcal{T}}\,\mathbf{u}$:

$$\text{proj}_{\mathcal{T}}\,\mathbf{u} \equiv \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j. \tag{10}$$

The idea that the best approximation of $\mathbf{u}$ in $\mathcal{T}$ is the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$ lends a welcome geometrical interpretation to the problem of best approximation. In fact, let us rephrase Theorem 9.10.1 in terms of orthogonal projection.

THEOREM 9.10.1′ *Best Approximation by Orthogonal Projection*
Let $\mathbf{u}$ be any vector in a normed inner product vector space $\mathcal{S}$ with natural norm

($\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$), and let $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ be an ON set in $\mathcal{S}$. Denote the subspace span$\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ of $\mathcal{S}$ as $\mathcal{T}$. Then the best approximation of $\mathbf{u}$ in $\mathcal{T}$ (i.e., of the form $c_1\hat{\mathbf{e}}_1 + \cdots + c_N\hat{\mathbf{e}}_N$) is given by the orthogonal projection of $\mathbf{u}$ onto $\mathcal{T}$, namely, by $\text{proj}_{\mathcal{T}}\,\mathbf{u}$.

9.10.2. Kronecker delta. When working with ON sets it is convenient to use the **Kronecker delta** symbol δ_{jk}, defined as

$$\delta_{jk} \equiv \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases} \tag{11}$$

and named after *Leopold Kronecker* (1823–1891), who contributed to algebra and the theory of equations. The subscripted j and k are usually positive integers. Clearly, δ_{jk} is symmetric in its indices j and k:

$$\delta_{jk} = \delta_{kj}. \tag{12}$$

To illustrate the use of the Kronecker delta, suppose that $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ is an ON basis for some space $\mathcal{S}$, and that we wish to expand a given $\mathbf{u}$ in $\mathcal{S}$ as

$$\mathbf{u} = \sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j. \tag{13}$$

To determine the c_j's, dot $\hat{\mathbf{e}}_k$ into both sides, where k is any integer such that $1 \leq k \leq N$, and use the fact that $\hat{\mathbf{e}}_j \cdot \hat{\mathbf{e}}_k = \delta_{jk}$ (because the $\hat{\mathbf{e}}_j$'s are ON):

$$\begin{aligned} \mathbf{u}\cdot\hat{\mathbf{e}}_k &= \left(\sum_{j=1}^{N} c_j\hat{\mathbf{e}}_j\right)\cdot\hat{\mathbf{e}}_k = \sum_{j=1}^{N} c_j\left(\hat{\mathbf{e}}_j\cdot\hat{\mathbf{e}}_k\right) \\ &= \sum_{j=1}^{N} c_j\delta_{jk} = c_k. \end{aligned} \tag{14}$$

Thus, $c_k = \mathbf{u}\cdot\hat{\mathbf{e}}_k$ for each $k = 1, 2, \dots, N$ so (13) becomes

$$\mathbf{u} = \sum_{j=1}^{N} \left(\mathbf{u}\cdot\hat{\mathbf{e}}_j\right)\hat{\mathbf{e}}_j. \tag{15}$$

Closure. Principal interest, in this brief section, is in the best approximation of a given vector $\mathbf{u}$ in a normed inner product vector space $\mathcal{S}$ in terms of an ON set $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ which falls short of being a basis for $\mathcal{S}$ inasmuch as $N < \dim \mathcal{S}$. Of course, if $N = \dim \mathcal{S}$ so the set is a basis, then we have the equality (15), but if $N < \dim \mathcal{S}$, then the best approximation of $\mathbf{u}$ is given by (5), best in the vector sense; that is, the norm of the error vector [i.e., the norm of the difference between the left- and right-hand sides of (15)] is minimized. From a geometric point of view, (15) says that the best approximation of $\mathbf{u}$ is the orthogonal projection of $\mathbf{u}$ onto the span of $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N$, which concept is explained in Examples 1 and 2, and which should be understood. In those examples we use the usual inner product for $\mathbb{R}^n$ (namely, $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + \cdots + u_nv_n$) and the corresponding natural norm (the Euclidean norm), but it should be understood that the results of this section (Theorems 9.10.1 and 9.10.1′) hold for any choice of the norm and inner product, provided that we use the "natural norm" $\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}$.

EXERCISES 9.10

1. We concluded from (4) that the best choice for the c_j's is $c_j = \alpha_j = \mathbf{u}\cdot\hat{\mathbf{e}}_j$. Show that this same result is obtained from (2) by setting $\partial\,\|\mathbf{E}\|^2/\partial c_j = 0$, and verify that the extremum thus obtained is a minimum.

2. Let $\mathcal{S}$ be $\mathbb{R}^5$, and let $N = 3$ with $\hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}(1,0,2,0,0)$, $\hat{\mathbf{e}}_2 = \frac{1}{\sqrt{6}}(2,0,-1,0,1)$, $\hat{\mathbf{e}}_3 = (0,0,0,1,0)$. Find the best approximation to the given $\mathbf{u}$ vector within span$\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$, and the norm of the error vector.

(a) $(3,-2,0,0,5)$ (b) $(0,0,0,2,1)$ (c) $(3,0,1,4,1)$
(d) $(1,1,0,1,1)$ (e) $(0,2,0,0,0)$ (f) $(1,0,-3,3,1)$
(g) $(0,7,0,3,0)$ (h) $(1,2,3,4,5)$ (i) $(5,4,3,2,1)$

3. Let $\mathcal{S}$ be $\mathbb{R}^4$, and let

$$\hat{\mathbf{e}}_1 = \tfrac{1}{\sqrt{3}}(1,1,0,-1), \quad \hat{\mathbf{e}}_2 = \tfrac{1}{\sqrt{3}}(1,-1,-1,0),$$
$$\hat{\mathbf{e}}_3 = \tfrac{1}{\sqrt{3}}(1,0,1,1), \quad \hat{\mathbf{e}}_4 = \tfrac{1}{\sqrt{3}}(0,1,-1,1).$$

Find the best approximation to $\mathbf{u} = (4,-2,1,6)$ within span$\{\hat{\mathbf{e}}_1\}$, span$\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$, span$\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$, and span$\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3, \hat{\mathbf{e}}_4\}$, and in each case compute the norm of the error vector, $\|\mathbf{E}\|$.

4. Same as Exercise 3, but for the given $\mathbf{u}$ vector.

(a) $(4,1,0,-1)$ (b) $(3,-1,1,2)$ (c) $(0,0,2,5)$
(d) $(1,2,4,4)$ (e) $(0,5,3,-1)$ (f) $(2,0,-1,-1)$

5. (*Bessel inequality*) Beginning with (4), derive the **Bessel inequality**

$$\boxed{\sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)^2 \le \|\mathbf{u}\|^2.} \tag{5.1}$$

Notice that if $\mathbf{u}$ happens to be in span$\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_k\}$, or if dim $\mathcal{S} = N$, then (5.1) becomes an equality. In two and three dimensions that equality is actually the Pythagorean theorem, and in more than three dimensions it amounts to an abstract extension of that theorem.

6. (*A different inner product*) In Examples 1 and 2 we use the "usual" inner product for $\mathbb{R}^n$, $\mathbf{u}\cdot\mathbf{v} = u_1v_1 + \cdots + u_nv_n$, but that is not the only acceptable one. In Example 3 of Section 9.6 we see that another acceptable inner product is

$$\mathbf{u}\cdot\mathbf{v} = w_1u_1v_1 + \cdots + w_nu_nv_n, \tag{6.1}$$

where the w_j's are fixed positive constants, or "weights."
(a) Rework Example 1 using the modified inner product $\mathbf{u}\cdot\mathbf{v} = 2u_1v_1 + u_2v_2$ and its corresponding natural norm $\|\mathbf{u}\| = \sqrt{2u_1^2 + u_2^2}$. Show that the resulting best approximation $\frac{29}{313}(12,5)$, which is not the same as the best approximation $\frac{17}{169}(12,5)$ given by (6). HINT: You will need to rescale $\mathbf{e}_1$ so as to be a unit vector according to the new norm.
(b) Whereas the error vector $\mathbf{AB}$ (Fig. 1) is orthogonal *and* perpendicular to span$\{\hat{\mathbf{e}}_1\}$, show that in this exercise the error vector is indeed orthogonal to span$\{\hat{\mathbf{e}}_1\}$, as promised in the text, but *not* perpendicular to it. To explain this "paradox," show that for the modified inner product the orthogonality of two nonzero vectors does not imply their perpendicularity.

7. Verify the last step in (14), that $\sum_{j=1}^{N} c_j\delta_{jk} = c_k$.

8. Verify the following, where (i,j,k,l) run from 1 to N.

(a) $\displaystyle\sum_{j=1} \delta_{ij} = 1$ (b) $\displaystyle\sum_{j} \delta_{ij}\delta_{jk} = \delta_{ik}$
(c) $\displaystyle\sum_{j}\sum_{k} \delta_{ij}\delta_{jk}\delta_{kl} = \delta_{il}$

Chapter 9 Review

We begin with the two- and three-dimensional "arrow vector" concept that is probably already familiar to you from an introductory course in physics, where the vectors denoted forces, velocities, and so on. For such vectors, vector addition $\mathbf{u}+\mathbf{v}$, scalar multiplication ($\alpha\mathbf{u}$), a zero vector ($\mathbf{0}$), a negative inverse $[-\mathbf{u} = (-1)\mathbf{u}]$, a norm ($\|\mathbf{u}\|$), a dot product

$$\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta, \tag{1}$$

and the angle $\theta = \cos^{-1}\left(\dfrac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}\right)$ between $\mathbf{u}$ and $\mathbf{v}$ are all defined.

From there, we generalize to abstract n-space, where $\mathbf{u} = (u_1, \dots, u_n)$, by defining vector addition, and so on, in such a way that they agree with the corresponding arrow vector definitions when $n = 2$ and $n = 3$. For instance,

$$\mathbf{u}\cdot\mathbf{v} = \sum_{j=1}^{n} u_j v_j, \tag{2}$$

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j^2}, \tag{3}$$

and

$$\theta = \cos^{-1}\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\,\|\mathbf{v}\|}. \tag{4}$$

From these definitions, we derived various *properties* such as

$$\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}, \qquad \text{(commutative)} \tag{5}$$

$$(\mathbf{u}+\mathbf{v})+\mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w}), \qquad \text{(associative)} \tag{6}$$

and so on, along with the following properties of the dot product and norm.

Dot Product

$$\textit{Commutative:}\qquad \mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}, \tag{7a}$$

$$\textit{Nonnegative:}\qquad \begin{aligned}\mathbf{u}\cdot\mathbf{u} &> 0 \quad \text{for all } \mathbf{u}\neq\mathbf{0}\\ &= 0 \quad \text{for } \mathbf{u}=\mathbf{0},\end{aligned} \tag{7b}$$

$$\textit{Linear:}\qquad (\alpha\mathbf{u}+\beta\mathbf{v})\cdot\mathbf{w} = \alpha(\mathbf{u}\cdot\mathbf{w})+\beta(\mathbf{v}\cdot\mathbf{w}), \tag{7c}$$

Norm

$$\textit{Scaling:}\qquad \|\alpha\mathbf{u}\| = |\alpha|\,\|\mathbf{u}\|, \tag{8a}$$

$$\textit{Nonnegative:}\qquad \begin{aligned}\|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u}\neq\mathbf{0}\\ &= 0 \quad \text{for } \mathbf{u}=\mathbf{0},\end{aligned} \tag{8b}$$

$$\textit{Triangular Inequality:}\qquad \|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\|+\|\mathbf{v}\|. \tag{8c}$$

To complete the extension to generalized vector space, we reverse the cart and the horse by elevating these various properties to the level of axioms, or requirements. That is, we let the fundamental objects, the vectors, be whatever we choose them to be, and then define addition and scalar multiplication operations, a zero vector, a negative inverse, a dot or "inner" product (if we wish), and a norm (if we wish), so that those axioms are satisfied. Our chief interest, in introducing generalized vector space, is in function spaces, but we will not work with function spaces until Chapter 17, when we study Fourier series and the Sturm–Liouville theory.

Next, we introduce the concept of *span* and *linear dependence*, primarily so that we can develop the idea of the *expansion* of a given vector in a vector space $\mathcal{S}$ in terms of a set of *base vectors* for $\mathcal{S}$. We define a set of vectors $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ to be a *basis* for $\mathcal{S}$ if each vector $\mathbf{u}$ in $\mathcal{S}$ can be expressed ("expanded") uniquely in the form $\mathbf{u} = \alpha_1\mathbf{e}_1 + \cdots + \alpha_k\mathbf{e}_k$, and prove that a set $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ is a basis for $\mathcal{S}$ if and only if it spans $\mathcal{S}$ and is LI (linearly independent). In particular, orthogonal bases are especially convenient because of the ease with which one can compute the expansion coefficients α_j. The result is

$$\mathbf{u} = \left(\frac{\mathbf{u}\cdot\mathbf{e}_1}{\mathbf{e}_1\cdot\mathbf{e}_1}\right)\mathbf{e}_1 + \cdots + \left(\frac{\mathbf{u}\cdot\mathbf{e}_k}{\mathbf{e}_k\cdot\mathbf{e}_k}\right)\mathbf{e}_k \tag{9}$$

if the basis is orthogonal, and

$$\mathbf{u} = (\mathbf{u}\cdot\hat{\mathbf{e}}_1)\,\hat{\mathbf{e}}_1 + \cdots + (\mathbf{u}\cdot\hat{\mathbf{e}}_k)\,\hat{\mathbf{e}}_k \tag{10}$$

if it is ON (orthonormal); (9) and (10) should be understood and remembered.

Finally, we study the question of the best *approximation* of a given vector $\mathbf{u}$ in a vector space $\mathcal{S}$ in terms of an ON set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ which falls short of being a basis for $\mathcal{S}$. We show that the best approximation (i.e., the one that minimizes the norm of the error vector) is

$$\mathbf{u} \approx \sum_{j=1}^{N} (\mathbf{u}\cdot\hat{\mathbf{e}}_j)\,\hat{\mathbf{e}}_j \tag{11}$$

which, in geometrical language, is the *orthogonal projection of* $\mathbf{u}$ *onto the span of* $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N$.

Chapter 10

Matrices and Linear Equations

10.1 Introduction

We have already met matrices in Section 8.3.3, but they were introduced there only as a notational convenience for the implementation of Gauss elimination and Gauss–Jordan reduction. In the present chapter we focus on matrix theory itself, which theory will enable us to obtain additional important results regarding the solution of systems of linear algebraic equations.

One way to view matrix theory is to think in terms of a parallel with function theory. In our mathematical training, we first study numbers – the points on a real number axis. Then we study functions, which are mappings, or transformations, from one real axis to another. For instance, $f(x) = x^2$ maps the point $x = 3$, say, on an x axis to the point $f = 9$ on an f axis. Just as functions act upon numbers, we shall see that matrices act upon vectors and are mappings from one vector space to another. Having studied vectors, in Chapter 9, we can now turn our attention to matrices.

Historically, matrix theory did not become a part of undergraduate engineering science curricula until around 1960, when digital computers became widely available in academia.

10.2 Matrices and Matrix Algebra

A **matrix** is a rectangular array of quantities that are called the **elements** of the matrix. Normally, the elements will be real numbers, although they may occasionally be other objects such as differential operators or even matrices. Some of these cases will be met as we go along; for the present, however, let us consider the elements to be *real numbers*. The complex case is studied in Chapter 12.

Specifically, any matrix $\mathbf{A}$ may be expressed as

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \tag{1}$$

where the brackets (or, in some texts, parentheses) are used to emphasize that the entire array is to be regarded as a single entity. A horizontal line of elements is called a **row**, and a vertical line is called a **column**. Counting rows from the top and columns from the left, then

$$a_{21} \quad a_{22} \quad \cdots \quad a_{2n} \qquad \text{and} \qquad \begin{matrix} a_{13} \\ a_{23} \\ \vdots \\ a_{m3}, \end{matrix}$$

say, are the second row and third column, respectively. Thus, we call the first subscript on a_{ij} the *row index*, and the second subscript the *column index*.

We usually use boldface capital letters to denote matrices and lightface lowercase letters to denote their elements. The matrix $\mathbf{A}$ in (1) is seen to have m rows and n columns and is therefore said to be $m \times n$ (read "m by n"); we shall refer to this as the *form* of $\mathbf{A}$. In some applications m and/or n may be infinite, but here we shall consider only matrices of finite size: $1 \leq m < \infty$, $1 \leq n < \infty$. Furthermore, m and n may, but need not, be equal. If $\mathbf{A}$ is small we may wish to dispense with the subscript notation for the elements. For example, if $m = n = 2$, we may prefer

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{to} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \tag{2}$$

but if $\mathbf{A}$ is large this becomes inconvenient. The double-subscript notation employed in (1) is especially convenient for digital computer calculations.

In view of the subscript notation in (1), one also writes

$$\mathbf{A} = \{a_{ij}\} \tag{3}$$

for short, where a_{ij} is called the *ij element* and $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Some authors write $a_{i,j}$ in place of a_{ij} to avoid ambiguity – for example, to prevent us from reading a_{21} as a-sub-twenty-one, but we will omit the commas, except when such ambiguity is not easily resolved from the context.

EXAMPLE 1. The matrices

$$\mathbf{A} = \begin{bmatrix} 3 & -1 \\ 0 & 2 \\ 7 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 8 & -6 & 0 \\ 1 & 3 & 27 \\ 2 & 9 & 4 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = [8, -7, 0, 4, 3]$$

are 3×2, 3×3, and 1×5, respectively. If we denote $\mathbf{B} = \{b_{ij}\}$, then $b_{11} = 8$, $b_{12} = -6$, $b_{32} = 9$, and so on. $\mathbf{C}$ happens to be a single row and it seems best to separate the elements by commas, but the commas are not essential. ▮

Two matrices are said to be **equal** if they are of the same form and if their corresponding elements are equal. For instance, none of the matrices above are equal, but if $\mathbf{D} = [8, -7, 0, 4, 3]$ then, $\mathbf{D} = \mathbf{C}$.

One may be tempted to identify $\mathbf{C}$, above, as a 5-tuple *vector* rather than as a 1×5 matrix. That would be a bit premature since vectors are not merely objects; they have rules for vector addition and scalar multiplication defined, whereas our matrices are, thus far, just mathematical "objects." In fact, our next step is to define some arithmetic operations for matrices so that they may be manipulated in useful ways. For vectors we defined two arithmetic operations, vector addition and scalar multiplication; for matrices we define three: matrix addition, scalar multiplication, *and* the multiplication of matrices.

Matrix addition. If $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ are any two matrices of the same form, say $m \times n$, then their sum $\mathbf{A} + \mathbf{B}$ is defined as

$$\boxed{\mathbf{A} + \mathbf{B} \equiv \{a_{ij} + b_{ij}\}} \tag{4}$$

and is itself an $m \times n$ matrix. If $\mathbf{A}$ and $\mathbf{B}$ are of the same form, they are said to be *conformable for addition*; if they are not of the same form, then $\mathbf{A} + \mathbf{B}$ is *not defined.*

EXAMPLE 2. If

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -6 \\ 1 & 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -1 & 2 & 0 \\ 15 & 6 & 3 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} 4 & 2 \\ -1 & 0 \end{bmatrix}, \tag{5}$$

then,

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 2 & -6 \\ 16 & 9 & 7 \end{bmatrix}, \tag{6}$$

but $\mathbf{A} + \mathbf{C}$ and $\mathbf{B} + \mathbf{C}$ are not defined since $\mathbf{A}$ and $\mathbf{B}$ are 2×3 while $\mathbf{C}$ is 2×2. ▮

Scalar multiplication. If $\mathbf{A} = \{a_{ij}\}$ is any $m \times n$ matrix and c is any scalar, their product is defined as

$$\boxed{c\mathbf{A} \equiv \{ca_{ij}\},} \tag{7}$$

and is itself an $m \times n$ matrix; we do not distinguish between $c\mathbf{A}$ and $\mathbf{A}c$. Furthermore, we denote

$$-\mathbf{A} \equiv (-1)\mathbf{A}. \tag{8}$$

In place of $\mathbf{A} + (-\mathbf{B})$, we simply write $\mathbf{A} - \mathbf{B}$, and call it the *difference* of $\mathbf{A}$ and $\mathbf{B}$, or $\mathbf{A}$ *minus* $\mathbf{B}$.

EXAMPLE 3. If A and C are the matrices in Example 1, then

$$3\mathrm{A} = \begin{bmatrix} 9 & -3 \\ 0 & 6 \\ 21 & 15 \end{bmatrix} \quad \text{and} \quad -\mathrm{C} = [-8, 7, 0, -4, -3]. \ \blacksquare$$

We shall list the important properties of matrix addition and scalar multiplication in a moment, but first let us define the so-called **zero matrix 0** to be any $m \times n$ matrix all the elements of which are zero. For example,

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{0} \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \mathbf{0},$$

the first being 2×3, the second being 3×1.

THEOREM 10.2.1 *Properties of Matrix Addition and Scalar Multiplication*
If $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are $m \times n$ matrices, $\mathbf{0}$ is an $m \times n$ zero matrix, and α, β are any scalars, then

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}, \qquad \text{(commutativity)} \tag{9a}$$
$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}), \qquad \text{(associativity)} \tag{9b}$$
$$\mathbf{A} + \mathbf{0} = \mathbf{A}, \tag{9c}$$
$$\mathbf{A} + (-\mathbf{A}) = \mathbf{0}, \tag{9d}$$
$$\alpha(\beta\mathbf{A}) = (\alpha\beta)\mathbf{A}, \qquad \text{(associativity)} \tag{9e}$$
$$(\alpha + \beta)\mathbf{A} = \alpha\mathbf{A} + \beta\mathbf{A}, \qquad \text{(distributivity)} \tag{9f}$$
$$\alpha(\mathbf{A} + \mathbf{B}) = \alpha\mathbf{A} + \alpha\mathbf{B}, \qquad \text{(distributivity)} \tag{9g}$$
$$1\mathbf{A} = \mathbf{A}, \tag{9h}$$
$$0\mathbf{A} = \mathbf{0}, \tag{9i}$$
$$\alpha\mathbf{0} = \mathbf{0}. \tag{9j}$$

The proof follows from the foregoing definitions and is left for the exercises.

Observe that there are no surprises in (9); the usual rules of arithmetic are seen to apply. For the special case where $\mathbf{A}$ consists of a single row (or column), we see that the definitions of addition and scalar multiplication above are identical to those introduced in Section 9.4 for n-tuple vectors. Thus, we may properly refer to the matrices

$$\mathbf{A} = [a_{11}, \dots, a_{1n}] \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} a_{11} \\ \vdots \\ a_{n1} \end{bmatrix} \tag{10}$$

as n-dimensional **row** and **column vectors**, respectively.

Matrix multiplication. Judging from the rather natural way in which matrix addition and scalar multiplication are defined, by (4) and (7), one might well expect the multiplication of two matrices $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{B} = \{b_{ij}\}$ to be defined only if $\mathbf{A}$ and $\mathbf{B}$ are of the same form, with the definition $\mathbf{AB} \equiv \{a_{ij}b_{ij}\}$. In fact, this is *not* the case. Instead, the standard definition of matrix multiplication is the one suggested by Cayley.[*] Called the *Cayley product*, it is as follows: if $\mathbf{A} = \{a_{ij}\}$ is any $m \times n$ matrix and $\mathbf{B} = \{b_{ij}\}$ is any $n \times p$ matrix (so that the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$), then the product $\mathbf{AB}$ is defined as

$$\boxed{\mathbf{AB} \equiv \left\{ \sum_{k=1}^{n} a_{ik}b_{kj} \right\}; \qquad (1 \le i \le m,\ 1 \le j \le p)} \tag{11}$$

that is, if we denote $\mathbf{AB} = \mathbf{C} = \{c_{ij}\}$, then

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}. \tag{12}$$

If the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$, then $\mathbf{A}$ and $\mathbf{B}$ are said to be *conformable for multiplication*; if not, the product $\mathbf{AB}$ is *not defined*. NOTE: The relative forms of $\mathbf{A}$, $\mathbf{B}$, and their product $\mathbf{C}$ are important and, as stated above, are as follows:

$$\begin{array}{ccccc} \mathbf{A} & \text{times} & \mathbf{B} & = & \mathbf{C}. \\ m \times n & & n \times p & & m \times p \\ \nwarrow & \text{equal} & \nearrow & & \end{array} \tag{13}$$

EXAMPLE 4. Suppose that

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -5 \\ 1 & 3 & 2 \\ 4 & 1 & -1 \\ 0 & 2 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 5 & 1 \\ -2 & 3 \\ 1 & 0 \end{bmatrix}.$$

[*] *Arthur Cayley* (1821–1895) produced around 200 papers in a 15-year period during which he was engaged in the practice of law. In 1863, he accepted a professorship of mathematics at Cambridge.

Then $\mathbf{A}$ is 4×3 and $\mathbf{B}$ is 3×2. Since the number of columns of $\mathbf{A}$ (namely, 3) is the same as the number of rows of $\mathbf{B}$, the product $\mathbf{AB}$ is defined and, according to (13), will be 4×2. According to the definition (12),

$$\mathbf{AB} = \underset{4\times 3}{\begin{bmatrix} 2 & 0 & -5 \\ 1 & 3 & 2 \\ \boxed{4} & \boxed{1} & -1 \\ 0 & 2 & 7 \end{bmatrix}} \underset{3\times 2}{\begin{bmatrix} 5 & \boxed{1} \\ -2 & \boxed{3} \\ 1 & 0 \end{bmatrix}} = \underset{4\times 2}{\begin{bmatrix} 5 & 2 \\ 1 & 10 \\ 17 & \boxed{7} \\ 3 & 6 \end{bmatrix}} = \mathbf{C}. \tag{14}$$

To compute c_{32}, for example, (12) gives

$$c_{32} = \sum_{k=1}^{3} \underset{\substack{\uparrow \\ \text{3rd row} \\ \text{of } \mathbf{A}}}{a_{3k}} \underset{\substack{\nwarrow \\ \text{2nd column} \\ \text{of } \mathbf{B}}}{b_{k2}} = (4)(1) + (1)(3) + (-1)(0) = 7,$$

as indicated by the arrows in (14). One more:

$$c_{11} = \sum_{k=1}^{3} a_{1k} b_{k1} = (2)(5) + (0)(-2) + (-5)(1) = 5.$$

We move across the rows of the first matrix and down the columns of the second.

COMMENT. Observe that c_{32} is the *dot product* of the third row of $\mathbf{A}$, considered as a 3-tuple vector, with the second column of $\mathbf{B}$. More generally, if $\mathbf{AB} = \mathbf{C} = \{c_{ij}\}$, then c_{ij} is the dot product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$. Thus, the number of elements in the rows of $\mathbf{A}$ (namely, the number of columns in $\mathbf{A}$) must equal the number of elements in the columns of $\mathbf{B}$ (namely, the number of rows of $\mathbf{B}$). ▮

EXAMPLE 5. Two more examples:

$$\overset{4\times 3}{\begin{bmatrix} 5 & 0 & -1 \\ 2 & 3 & 4 \\ 1 & 0 & 6 \\ 0 & 0 & 1 \end{bmatrix}} \overset{3\times 1}{\begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}} = \overset{4\times 1}{\begin{bmatrix} (5)(3) + (0)(2) + (-1)(5) \\ (2)(3) + (3)(2) + (4)(5) \\ (1)(3) + (0)(2) + (6)(5) \\ (0)(3) + (0)(2) + (1)(5) \end{bmatrix}} = \begin{bmatrix} 10 \\ 32 \\ 33 \\ 5 \end{bmatrix},$$

and

$$\overset{2\times 2}{\begin{bmatrix} -3 & 1 \\ 10 & 2 \end{bmatrix}} \overset{2\times 2}{\begin{bmatrix} 1 & 0 \\ 2 & 4 \end{bmatrix}} = \overset{2\times 2}{\begin{bmatrix} (-3)(1) + (1)(2) & (-3)(0) + (1)(4) \\ (10)(1) + (2)(2) & (10)(0) + (2)(4) \end{bmatrix}}$$

$$= \begin{bmatrix} -1 & 4 \\ 14 & 8 \end{bmatrix}. \quad ▮$$

It is extremely important to see that *matrix multiplication is not, in general, commutative; that is,*

$$\mathbf{AB} \neq \mathbf{BA}, \tag{15}$$

except in exceptional cases. For suppose that $\mathbf{A} = m \times n$ and $\mathbf{B} = n \times p$ (i.e., $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$) so that $\mathbf{AB}$ is at least defined. However, $\mathbf{BA} = (n \times p)(m \times n)$ is not even defined, let alone equal to $\mathbf{AB}$, unless $p = m$. Assuming that that is the case,

$$\mathbf{BA} = (n \times m)(m \times n) = n \times n, \tag{16a}$$

whereas

$$\mathbf{AB} = (m \times n)(n \times m) = m \times m. \tag{16b}$$

Comparing (16a) with (16b), we see that we must also have $m = n$ if $\mathbf{AB}$ and $\mathbf{BA}$ are to be of the same form and hence *possibly* equal. Thus, a necessary condition for $\mathbf{AB}$ to equal $\mathbf{BA}$ (i.e., for $\mathbf{A}$ and $\mathbf{B}$ to commute under multiplication) is that $\mathbf{A}$ and $\mathbf{B}$ both be $n \times n$.

EXAMPLE 6. If

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 3 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -2 & 3 \\ 3 & -4 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 2 & 3 \\ 1 & 0 \end{bmatrix},$$

we find that $\mathbf{A}$ and $\mathbf{B}$ commute ($\mathbf{AB} = \mathbf{BA}$), but $\mathbf{A}$ and $\mathbf{C}$ do not ($\mathbf{AC} \neq \mathbf{CA}$), nor do $\mathbf{B}$ and $\mathbf{C}$. ▮

Thus, the condition that $\mathbf{A}$ and $\mathbf{B}$ must both be $n \times n$, for $\mathbf{A}$ and $\mathbf{B}$ to commute, is *necessary but not sufficient.* In view of the importance of which factor is first and which is second in a matrix product $\mathbf{AB}$, we sometimes say that $\mathbf{B}$ is **pre-multiplied** by $\mathbf{A}$, and $\mathbf{A}$ is **post-multiplied** by $\mathbf{B}$ so as to leave no doubt as to which factor is first and which is second.

The lack of commutativity, in general, is a major setback so we must wonder why Cayley's definition has been adopted rather than the simpler one that comes to mind, $\mathbf{AB} \equiv \{a_{ij}b_{ij}\}$, which would surely yield commutativity since $\mathbf{BA} = \{b_{ij}a_{ij}\} = \{a_{ij}b_{ij}\} = \mathbf{AB}$ (the second equality following from the commutativity of the multiplication of ordinary numbers). A sufficiently compelling reason to use Cayley's definition involves the application of matrix notation to systems of linear algebraic equations for it turns out that, with Cayley's definition of multiplication, the system of m linear algebraic equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m \end{aligned} \tag{17}$$

in the n unknowns $x_1, \ldots, x_n$ is equivalent to the single compact matrix equation

$$\mathbf{Ax} = \mathbf{c}, \tag{18}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}. \tag{19}$$

A is called the **coefficient matrix**, and **x** is the unknown; that is, its components are the unknowns $x_1, \ldots, x_n$. To verify the claimed equivalence, work out the product **Ax** , and set the result equal to **c**. That step gives

$$\underset{m \times 1}{\begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ a_{21}x_1 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n \end{bmatrix}} = \underset{m \times 1}{\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix}}. \tag{20}$$

These two $m \times 1$ matrices (or m-dimensional column vectors) will be equal if and only if each of their corresponding m elements (or components) are equal. Thus, (18) is equivalent to the m scalar equations (17), so (17) and (18) are equivalent, as claimed. This important result is of course a consequence of (11) and provides strong support for adopting that definition of matrix multiplication.

Any $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is said to be **square**, and of **order** n,* and the elements $a_{11}, a_{22}, \ldots, a_{nn}$ are said to lie on the **main diagonal** of **A** – that is, the diagonal from the upper left corner to the lower right corner. Notice that to be able to multiply any matrix **A** with itself, **A** needs to be square. For suppose that **A** is $m \times n$; then we have

$$\underset{m \times \underline{n}}{\mathbf{A}} \qquad \underset{\underline{m} \times n}{\mathbf{A}}$$

and we need n (the number of columns in the first matrix) to equal m (the number of rows in the second) for the multiplication to be defined. If **A** *is* square and p is any positive integer, we define

$$\underbrace{\mathbf{A}\mathbf{A}\cdots\mathbf{A}}_{p \text{ factors}} \equiv \mathbf{A}^p. \tag{21}$$

The familiar laws of exponents,

$$\mathbf{A}^p\mathbf{A}^q = \mathbf{A}^{p+q}, \qquad (\mathbf{A}^p)^q = \mathbf{A}^{pq} \tag{22}$$

follow for any positive integers p and q.

*Thus, we distinguish between form and order: the **form** of an $m \times n$ matrix is $m \times n$, the **order** of an $n \times n$ matrix is n.

If, in particular, the only nonzero elements of a square matrix lie on the main diagonal, $\mathbf{A}$ is said to be a **diagonal matrix**. For example,

$$\mathbf{A} = \begin{bmatrix} 3 & 0 \\ 0 & -5 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 7 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

are diagonal, and any diagonal matrix of order n can be denoted as

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_{nn} \end{bmatrix}, \tag{23}$$

where not all of the d_{jj}'s are zero,* and all of the off-diagonal elements are zero; that is, $d_{ij} = 0$ if $i \neq j$. It is left for the exercises to show that

$$\mathbf{D}^p = \begin{bmatrix} d_{11}^p & 0 & \cdots & 0 \\ 0 & d_{22}^p & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_{nn}^p \end{bmatrix}, \tag{24}$$

for any positive integer p.

If, furthermore, $d_{11} = d_{22} = \cdots = d_{nn} = 1$, then $\mathbf{D}$ is called the **identity matrix I**. Thus,

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & 1 \end{bmatrix} = \{\delta_{ij}\}, \tag{25}$$

where δ_{ij} is the **Kronecker delta** symbol defined in Section 9.10.2, namely,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{26}$$

It is sometimes convenient to include a subscript n to indicate the order of $\mathbf{I}$. For example,

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

*If all of the diagonal elements were zero as well, then *all* the elements would be zero, and it would be more reasonable to describe $\mathbf{D}$ as a zero matrix, $\mathbf{0}$.

The key property of the identity matrix is that if $\mathbf{A}$ is any square matrix of the same order as $\mathbf{I}$, then

$$\mathbf{IA} = \mathbf{AI} = \mathbf{A}, \tag{27}$$

proof of which is left for the exercises. In other words, $\mathbf{I}$ is the matrix analog of the number 1 in scalar arithmetic! From the first equality in (27), we see that one case in which commutativity *does* hold is when one of the matrices is the identity matrix $\mathbf{I}$.

Finally, it is convenient to extend our definition of $\mathbf{A}^p$ [recall (21)] to the case where $p = 0$. If $\mathbf{A}$ is any $n \times n$ matrix, we define

$$\mathbf{A}^0 \equiv \mathbf{I}, \tag{28}$$

where $\mathbf{I}$ is an $n \times n$ identity matrix.

Perhaps we should take a moment to mention that whereas the identity matrix $\mathbf{I}$ is necessarily square, the zero matrix $\mathbf{0}$ is simply $m \times n$, not necessarily square. It is readily verified that

$$\mathbf{0A} = \mathbf{0} \quad \text{and} \quad \mathbf{A0} = \mathbf{0} \tag{29}$$

for any matrix $\mathbf{A}$. In the first of these equations, suppose that the $\mathbf{0}$ on the left is $m \times n$ and that $\mathbf{A}$ is $n \times p$. Then the $\mathbf{0}$ on the right is $m \times p$; that is, it is not necessarily of the same form as the one on the left.

In view of the general failure of commutativity, as stated in (15), we may well wonder if any other familiar arithmetic rules fail to hold for the multiplication of matrices. The answer is yes; the following rules for real numbers (a, b, c) do *not* carry over to matrices:

1. $ab = ba$ (commutativity).
2. If $ab = ac$ and $a \neq 0$, then $b = c$ (cancellation rule).
3. If $ab = 0$, then $a = 0$ and/or $b = 0$.
4. If $a^2 = 1$, then $a = +1$ or -1.

To add emphasis, we state these difficulties as a theorem.

THEOREM 10.2.2 *"Exceptional" Properties of Matrix Multiplication*

(i) $\mathbf{AB} \neq \mathbf{BA}$ in general.
(ii) Even if $\mathbf{A} \neq \mathbf{0}$, $\mathbf{AB} = \mathbf{AC}$ does not imply that $\mathbf{B} = \mathbf{C}$.
(iii) $\mathbf{AB} = \mathbf{0}$ does not imply that $\mathbf{A} = \mathbf{0}$ and/or $\mathbf{B} = \mathbf{0}$.
(iv) $\mathbf{A}^2 = \mathbf{I}$ does not imply that $\mathbf{A} = +\mathbf{I}$ or $-\mathbf{I}$.

The first of these has already been discussed, and the others are discussed in the exercises. Theorem 10.2.2 notwithstanding, several important properties do carry over from the multiplication of real numbers to the multiplication of matrices:

THEOREM 10.2.3 *"Ordinary" Properties of Matrix Multiplication*
If α, β are scalars, and the matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are suitably conformable, then

$$\begin{aligned}(\alpha\mathbf{A})\mathbf{B} &= \mathbf{A}(\alpha\mathbf{B}) = \alpha(\mathbf{AB}), && \text{(associativity)} && \text{(30a)}\\ \mathbf{A}(\mathbf{BC}) &= (\mathbf{AB})\mathbf{C}, && \text{(associativity)} && \text{(30b)}\\ (\mathbf{A}+\mathbf{B})\mathbf{C} &= \mathbf{AC}+\mathbf{BC}, && \text{(distributivity)} && \text{(30c)}\\ \mathbf{C}(\mathbf{A}+\mathbf{B}) &= \mathbf{CA}+\mathbf{CB}, && \text{(distributivity)} && \text{(30d)}\\ \mathbf{A}(\alpha\mathbf{B}+\beta\mathbf{C}) &= \alpha\mathbf{AB}+\beta\mathbf{AC}. && \text{(linearity)} && \text{(30e)}\end{aligned}$$

Proof is left for the exercises.

Partitioning. Let us close this section with a discussion of the **partitioning** of matrices. The idea is that any matrix $\mathbf{A}$ (which is larger that 1×1) may be partitioned into a number of smaller matrices called **blocks** by vertical lines that extend from bottom to top, and horizontal lines that extend from left to right.

EXAMPLE 7. *Partitioning.*

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -3\\ 5 & 2 & 7\\ 1 & 3 & 0\\ 0 & 4 & 6\end{bmatrix} = \left[\begin{array}{cc|c} 2 & 0 & -3\\ 5 & 2 & 7\\ \hline 1 & 3 & 0\\ \hline 0 & 4 & 6\end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\\ \mathbf{A}_{31} & \mathbf{A}_{32}\end{bmatrix}, \tag{31}$$

where the blocks are

$$\mathbf{A}_{11} = \begin{bmatrix} 2 & 0\\ 5 & 2\end{bmatrix}, \quad \mathbf{A}_{12} = \begin{bmatrix} -3\\ 7\end{bmatrix}, \quad \mathbf{A}_{21} = [1, 3] \quad \mathbf{A}_{22} = [0],$$

and so on. Clearly, the partition is not unique. In the present example we could also have set

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -3\\ 5 & 2 & 7\\ 1 & 3 & 0\\ 0 & 4 & 6\end{bmatrix} = \left[\begin{array}{c|c|c} 2 & 0 & -3\\ 5 & 2 & 7\\ 1 & 3 & 0\\ 0 & 4 & 6\end{array}\right] = [\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{13}], \tag{32}$$

say. ▮

While the matrices used here as illustrations are kept small for convenience, those encountered in modern applications may be quite large, for example 600 × 800. Even with modern computers such large matrices create special computational problems, and it is often advantageous to work instead with a number of smaller matrices through the use of partitioning. Such advantages might well prove illusory, however, were it not for the fact that the usual matrix arithmetic *can* be carried out with partitioned matrices.

Specifically, if $\mathbf{A}$ and $\mathbf{B}$ are partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1n} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \cdots & \mathbf{A}_{mn} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1q} \\ \vdots & \vdots & & \vdots \\ \mathbf{B}_{p1} & \mathbf{B}_{p2} & \cdots & \mathbf{B}_{pq} \end{bmatrix}, \tag{33}$$

then

$$\alpha\mathbf{A} = \begin{bmatrix} \alpha\mathbf{A}_{11} & \alpha\mathbf{A}_{12} & \cdots & \alpha\mathbf{A}_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha\mathbf{A}_{m1} & \alpha\mathbf{A}_{m2} & \cdots & \alpha\mathbf{A}_{mn} \end{bmatrix}; \tag{34}$$

if $m = p$ and $n = q$ and each $\mathbf{A}_{ij}$ block is of the same form as the corresponding $\mathbf{B}_{ij}$ block, then

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{B}_{11} & \mathbf{A}_{12} + \mathbf{B}_{12} & \cdots & \mathbf{A}_{1n} + \mathbf{B}_{1n} \\ \vdots & \vdots & & \vdots \\ \mathbf{A}_{m1} + \mathbf{B}_{m1} & \mathbf{A}_{m2} + \mathbf{B}_{m2} & \cdots & \mathbf{A}_{mn} + \mathbf{B}_{mn} \end{bmatrix}; \tag{35}$$

and if $n = p$ and we denote $\mathbf{AB} = \mathbf{C}$, then

$$\mathbf{C}_{ij} = \sum_{k=1}^{n} \mathbf{A}_{ik}\mathbf{B}_{kj}, \tag{36}$$

provided that the number of columns in each $\mathbf{A}_{ik}$ is the same as the number of rows in the corresponding $\mathbf{B}_{kj}$, so that the products in (36) are defined.

Verification of these three claims, (34) to (36), is left for the exercises.

EXAMPLE 8. If

$$\mathbf{A} = \left[\begin{array}{cc|c} 2 & 4 & 1 \\ -1 & 3 & 0 \\ \hline 5 & 4 & 6 \end{array}\right] = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{37}$$

and

$$\mathbf{B} = \left[\begin{array}{cc|c} 0 & 1 & 3 \\ 2 & -4 & -1 \\ \hline 5 & 8 & 2 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{38}$$

then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}.$$

Working out the the elements,

$$\mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} = \begin{bmatrix} 2 & 4 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & -4 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} [5, 8]$$

$$\begin{aligned}
&= \begin{bmatrix} 8 & -14 \\ 6 & -13 \end{bmatrix} + \begin{bmatrix} 5 & 8 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 13 & -6 \\ 6 & -13 \end{bmatrix}, \\
\mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} &= \begin{bmatrix} 2 & 4 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} [2] \\
&= \begin{bmatrix} 2 \\ -6 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ -6 \end{bmatrix}, \\
\mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} &= [5, 4] \begin{bmatrix} 0 & 1 \\ 2 & -4 \end{bmatrix} + [6][5, 8] \\
&= [8, -11] + [30, 48] = [38, 37],
\end{aligned}$$

and

$$\mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} = [5, 4] \begin{bmatrix} 3 \\ -1 \end{bmatrix} + [6][2] = [23],$$

so

$$\mathbf{AB} = \left[\begin{array}{cc|c} 13 & -6 & 4 \\ 6 & -13 & -6 \\ \hline 38 & 37 & 23 \end{array}\right] = \begin{bmatrix} 13 & -6 & 4 \\ 6 & -13 & -6 \\ 38 & 37 & 23 \end{bmatrix}, \tag{39}$$

which is the same result as obtained by the multiplication of the unpartitioned matrices $\mathbf{A}$ and $\mathbf{B}$.

COMMENT 1. By no means do we claim that partitioning made the preceding calculation easier; our aim was simply to illustrate the idea of partitioning.

COMMENT 2. The partition

$$\mathbf{B} = \left[\begin{array}{c|cc} 0 & 1 & 3 \\ 2 & -4 & -1 \\ \hline 5 & 8 & 2 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{40}$$

in place of (38), would also be conformable with (37) for the multiplication $\mathbf{AB}$ and would lead to the same result, (39). On the other hand, neither of the partitions

$$\mathbf{B} = \left[\begin{array}{c|cc} 0 & 1 & 3 \\ \hline 2 & -4 & -1 \\ 5 & 8 & 2 \end{array}\right] \quad \text{or} \quad \mathbf{B} = \left[\begin{array}{cc|c} 0 & 1 & 3 \\ \hline 2 & -4 & -1 \\ 5 & 8 & 2 \end{array}\right] \tag{41}$$

would be conformable with (37) for the product $\mathbf{AB}$. For example, the term $\mathbf{A}_{11}\mathbf{B}_{11}$ would not be defined then since $\mathbf{A}_{11}$ has two columns whereas the $\mathbf{B}_{11}$'s in (41) have only one row. ▮

Let us close by giving two results, for reference, that will be used later on. Both use partitioning to work out the product of two matrices, $\mathbf{A}$ and $\mathbf{B}$. First, if we partition $\mathbf{B}$ into columns $\mathbf{c}_1, \ldots, \mathbf{c}_n$, then

$$\mathbf{AB} = \mathbf{A}\left[\mathbf{c}_1, \ldots, \mathbf{c}_n\right] = \left[\mathbf{Ac}_1, \ldots, \mathbf{Ac}_n\right]. \tag{42}$$

Second, if we also partition $\mathbf{A}$ into rows $\mathbf{r}_1, \dots, \mathbf{r}_m$, then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_m \end{bmatrix} [\mathbf{c}_1, \dots, \mathbf{c}_n] = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{c}_1 & \cdots & \mathbf{r}_1 \cdot \mathbf{c}_n \\ \vdots & & \vdots \\ \mathbf{r}_m \cdot \mathbf{c}_1 & \cdots & \mathbf{r}_m \cdot \mathbf{c}_n \end{bmatrix}. \tag{43}$$

That is, the i, j element of $\mathbf{AB}$ is $\mathbf{r}_i$ dotted with $\mathbf{c}_j$.

Closure. We define matrices and three arithmetic operations or matrices: addition, multiplication by a scalar, and multiplication. Subtraction is accounted for by addition and multiplication by a scalar: $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-1)\mathbf{B}$. But no division operation is defined for matrices.

It is emphasized that multiplication is not commutative (i.e., $\mathbf{AB} \neq \mathbf{BA}$ in general). This "failure" and several others are listed in Theorem 10.2.2. It is suggested that these shortcomings of the Cayley definition of matrix multiplication are more than offset by the fact that it permits us to express a system of m linear algebraic equations in the n unknowns $x_1, \dots, x_n$ in compact matrix form as $\mathbf{Ax} = \mathbf{c}$.

Computer software. As mentioned in Section 8.3, the *Maple* system contains many linear algebra commands within the linalg package, among which **evalm** is especially useful. For instance, to evaluate $(\mathbf{AB})^2\mathbf{P}^3 - 5\mathbf{Q}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix},$$

first enter

```
with(linalg):
```

and return. Then enter

$$A := \text{array}\ ([[1, 2], [3, 0]]): \tag{44}$$

and return. (If you wish the $\mathbf{A}$ matrix to be printed, type a semicolon in place of the final colon.) Similarly, for

$$B := \text{array}\ ([[1, -1], [0, 2]]): \tag{45}$$
$$P := \text{array}\ ([[1, 1], [1, 1]]): \tag{46}$$
$$Q := \text{array}\ ([[1, -1], [2, 3]]): \tag{47}$$

Then enter

```
evalm ((A &* B)^2 &* P^3 - 5 * Q);
```

and return. The printed output is the result

$$\begin{bmatrix} 11 & 21 \\ 38 & 33 \end{bmatrix}$$

Note that matrix multiplication of **A** and **B** is denoted as $A\&*B$ (not $\mathbf{A}*\mathbf{B}$), exponentiation to a positive integer power is denoted with ^, and multiplication of a matrix by a scalar is denoted by $*$. If we want $((\mathbf{AB})^2\mathbf{P}^3 - 5\mathbf{Q})^4$, we can use a quotation mark to carry the matrix $\begin{bmatrix} 11 & 21 \\ 38 & 33 \end{bmatrix}$ forward. Thus, enter

```
evalm(”^4);
```

and return. The result is

$$\begin{bmatrix} 2389489 & 2592744 \\ 4691632 & 5105697 \end{bmatrix}$$

Alternative to the array format indicated above, we can use a matrix format. For instance,

$$A := \text{matrix}\ (2, 2,\ [1, 2, 3, 0]):$$

is equivalent to the array format shown above, where the "2, 2" denotes that **A** is a 2×2 matrix.

EXERCISES 10.2

1. Given the matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 3 \\ 2 & -5 \\ 1 & 10 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & -1 \\ 0 & 2 \end{bmatrix},$$

$$\mathbf{x} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad \mathbf{y} = [-1, 2],$$

work out whichever of the products **AB**, **BA**, **Ax**, **xA**, **Bx**, **xB**, **yB**, $\mathbf{A}^2$, $\mathbf{B}^2$, $\mathbf{x}^2$, **xy**, and **yx** are defined.

2. Let **A** be 6×4, **B** be 4×4, **C** be 4×3, **D** and **E** be 3×1. Determine which of the following are defined, and for those that are, give the form of the resulting matrix.

(a) $\mathbf{A}^{10}$ (b) $\mathbf{B}^{10}$
(c) **ABC** (d) **ABCD**
(e) **ACBD** (f) **CD** + **E**
(g) **C**(2**D** − **E**) (h) **AB** + **AC**
(i) **BC** + **CB** (j) 3**BA** − 5**CD**

3. Evaluate the products

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \text{and}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

4. Suppose that **A** is $n \times n$, **x** is $n \times 1$, and c is a scalar. Can we re-express $\mathbf{Ax} = c\mathbf{x}$ as $(\mathbf{A} - c)\mathbf{x} = 0$? Explain.

5. If **A** and **B** are square matrices of the same order, are the following correct? Explain.

(a) $(\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$
(b) $(\mathbf{A} + \mathbf{B})(\mathbf{A} - \mathbf{B}) = \mathbf{A}^2 - \mathbf{B}^2$
(c) $(\mathbf{AB})^2 = \mathbf{A}^2\mathbf{B}^2$
(d) $(\mathbf{AB})^3 = \mathbf{A}^3\mathbf{B}^3$

6. (a) If p is a positive integer, does **A** need to be square for $\mathbf{A}^p$ to be defined? Explain.
(b) Let **A** be $m \times n$ and **B** be $p \times q$. What restrictions, if any, need must be satisfied by m, n, p, q if $(\mathbf{AB})^2$ is to exist (i.e., be defined)?

7. Expand each of the following [e.g., the "expanded" version of (**A** + **B**)**C** would be **AC** + **BC**], assuming that all of the matrices are suitably conformable. Justify each step by citing the relevant equation number in Theorem 10.2.1 or 10.2.3.

(a) (2**A** + **B**)(**A** + 2**B**) (b) (**A** + **B**)**C**(**D** + **E**)
(c) $(\mathbf{A} + \mathbf{B})^3$ (d) (**A** − 3**I**)(2**A** + **I**)

8. Given $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$, and $\mathbf{D} = \begin{bmatrix} 0 & 5 & 7 \\ 0 & 0 & 8 \\ 0 & 0 & 0 \end{bmatrix}$, evaluate each of the following.

(a) $\mathbf{A}^{100}$ (b) $\mathbf{B}^{100}$
(c) $\mathbf{C}^{100}$ (d) $\mathbf{D}^{100}$
(e) $(\mathbf{ABC})^3$ (f) $(\mathbf{CBA})^3$
(g) $\mathbf{B}^4\mathbf{C}^4$ and $(\mathbf{BC})^4$ (h) $\mathbf{C}^3\mathbf{B}^3$ and $(\mathbf{CB})^3$

9. Any diagonal matrix whose diagonal elements are all equal is called a **scalar matrix**. If

$$\mathbf{S} = \begin{bmatrix} k & 0 & \cdots & 0 \\ 0 & k & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & k \end{bmatrix}$$

is $n \times n$, and $\mathbf{A}$ is any $n \times n$ matrix, show that

$$\mathbf{AS} = \mathbf{SA} = k\mathbf{A}.$$

What can be said if, instead, $\mathbf{A}$ is $m \times n$ $(m \neq n)$?

10. If for any given vector $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_4 \end{bmatrix}$, the product $\mathbf{Ax}$ is the column vector given below, find $\mathbf{A}$.

(a) $\begin{bmatrix} x_1 - 3x_4 \end{bmatrix}$ (b) $\begin{bmatrix} x_2 + x_3 - x_4 \\ x_1 + 5x_3 \end{bmatrix}$

(c) $\begin{bmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_4 \end{bmatrix}$ (d) $\begin{bmatrix} 2x_1 - x_3 - x_4 \\ -2x_1 + x_2 \\ x_2 + x_4 \end{bmatrix}$

(e) $\begin{bmatrix} x_4 \\ x_3 \\ x_2 \\ x_1 \end{bmatrix}$ (f) $\begin{bmatrix} x_1 + 3x_4 \\ x_2 - x_4 \\ x_3 + x_4 \\ x_4 \\ x_3 - 2x_1 \end{bmatrix}$

11. Make up a specific pair of matrices, $\mathbf{A}$ and $\mathbf{B}$, both nonzero, such that $\mathbf{AB} = \mathbf{0}$, where

(a) $\mathbf{A}$ is 2×2 and $\mathbf{B}$ is 2×2
(b) $\mathbf{A}$ is 5×2 and $\mathbf{B}$ is 2×2
(c) $\mathbf{A}$ is 1×2 and $\mathbf{B}$ is 2×4
(d) $\mathbf{A}$ is 4×3 and $\mathbf{B}$ is 3×2

12. Given the partitioned matrices $\mathbf{A}$ and $\mathbf{B}$, below, carry out the products $\mathbf{A}^2$ and $\mathbf{AB}$ for those cases in which the partitioning is suitable, i.e., conformable. If the partitioning is not suitable, explain why it is not.

(a) $\mathbf{A} = \left[\begin{array}{cc|c} 2 & 0 & -1 \\ 1 & -1 & 0 \\ \hline 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{cc|c} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

(b) $\mathbf{A} = \left[\begin{array}{cc|c} 2 & 0 & -1 \\ 1 & -1 & 0 \\ \hline 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{c|cc} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

(c) $\mathbf{A} = \left[\begin{array}{c|cc} 2 & 0 & -1 \\ \hline 1 & -1 & 0 \\ 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{cc|c} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

13. Show that $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n$ can be expressed in the form

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

14. (a) If two unpartitioned matrices are not conformable for addition, can they be rendered conformable by suitable partitioning? Explain.
(b) Same as part (a), but for multiplication.

15. If there is some positive integer p such that $\mathbf{A}^p = \mathbf{0}$, then $\mathbf{A}$ is said to be **nilpotent** (mnemonic: "potentially nil"). A square matrix $\mathbf{A} = \{a_{ij}\}$ such that $a_{ij} = 0$ for all $i > j$ is said to be an **upper triangular matrix**; if $a_{ij} = 0$ for all $i < j$, then $\mathbf{A}$ is a **lower triangular matrix**. A matrix is said to be a **triangular** if it is either upper triangular or lower triangular.

(a) Every upper triangular matrix with null main diagonal (so that $a_{ij} = 0$ for all $i \geq j$) is nilpotent. Verify this result for second-, third-, and fourth-order matrices.
(b) In fact, show that *every* upper triangular matrix (of finite order) with null main diagonal is nilpotent. HINT: Use partitioning and induction.
(c) Is every lower diagonal matrix with null main diagonal nilpotent? Explain.
(d) If $\mathbf{A}^p = \mathbf{0}$, show that $(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{p-1})(\mathbf{I}-\mathbf{A}) = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{p-1}) = \mathbf{I}$.

16. If $\mathbf{A}^2 = \mathbf{I}$, then $\mathbf{A}$ is called **involutory.**

(a) Show, using Theorems 10.2.1 to 10.2.3 and (27), that $\mathbf{A}$ is involutory if and only if

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A}) = \mathbf{0}.$$

(b) Give an example of an involutory matrix other than $\mathbf{I}$ and $-\mathbf{I}$. Thus, observe that $\mathbf{A}^2 = \mathbf{I}$ does *not* imply that $\mathbf{A} = \pm\mathbf{I}$.
(c) Determine the most general 2×2 matrix that is involutory.

17. (a) Prove (9a) to (9c).
(b) Prove (9d) to (9f).
(c) Prove (9g) to (9i).

18. In Theorem 10.2.2, prove

(a) (i) (b) (ii) (c) (iii) (d) (iv)

19. (a) Verify (24).
(b) Verify (27).
(c) Verify (30a) and (30b).
(d) Verify (30c) and (30d).
(e) Show that (30e) follows from (30a)–(30d).

20. Show that the most general matrix that commutes with $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ is $\begin{bmatrix} \alpha-\beta & 2\beta/3 \\ \beta & \alpha \end{bmatrix}$, where α and β are arbitrary.

21. Given $\mathbf{A}$, find the most general matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{0}$.

(a) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (b) $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 0 & 0 \end{bmatrix}$

(c) $\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 5 & 0 \end{bmatrix}$ (d) $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix}$

22. Explore and discuss the advantages and disadvantages of defining $c_{ij} = a_{ij}b_{ij}$ in place of the Cayley product (13).

23. (*Transition probability matrix*) Selling their valuables, professors A and B raise \$2 apiece, and proceed to match coins at \$1 per match. There arise five possible states:

$$\begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5 \\ 04 & 13 & 22 & 31 & 40 \end{array}$$

In state S_2, for example, A has \$1 and B has \$3. If either player is bankrupted (A is bankrupt in state S_1, B in state S_5), the game is over. Let $p_{ij}^{(n)}$ be the n-step transition probability, i.e., the probability of changing from state S_i to state S_j in n matches. For $n = 1$ we see that

$$\mathbf{P}^{(1)} = \{p_{ij}^{(1)}\} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

For example, beginning in S_2, say, one match will necessarily move us to S_1 or S_3, with 50% probability in each case; thus $p_{21}^{(1)} = p_{23}^{(1)} = \frac{1}{2}$. Precisely \$1 will change hands so that $p_{22}^{(1)} = p_{24}^{(1)} = p_{25}^{(1)} = 0$. Further, once in S_1 (or S_5) we remain there (according to the rules), so $p_{11}^{(1)} = 1$, $p_{12}^{(1)} = p_{13}^{(1)} = p_{14}^{(1)} = p_{15}^{(1)} = 0$. Show, by any convincing arguments or discussion, that

$$p_{ij}^{(2)} = \sum_{k=1}^{5} p_{ik}^{(1)}p_{kj}^{(1)}, \quad p_{ij}^{(3)} = \sum_{k=1}^{5} p_{ik}^{(2)}p_{kj}^{(1)}, \quad \text{etc.}$$

or, in matrix notation, $\mathbf{P}^{(2)} = [\mathbf{P}^{(1)}]^2$, $\mathbf{P}^{(3)} = [\mathbf{P}^{(1)}]^3$, etc. Use this result to determine $\mathbf{P}^{(2)}$ and $\mathbf{P}^{(3)}$. What is the probability that A is bankrupt after (at most) three matches if A starts with \$2? \$3? \$1? NOTE: $\mathbf{P}^{(1)}$ is an example of a **Markov matrix**. We meet Markov matrices again in Chapter 11.

24. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 3 & 0 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 3 & 25 \\ 2 & 0.1 & -6 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 9 & 1 & -1 \\ 2 & 0 & 7 \\ 0 & 4 & 6 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 6 & 5 \\ 4 & 1 \end{bmatrix}.$$

NOTE: The letters D, E, I, O, S, and W are "protected," in *Maple*, for other purposes. The problem: use computer software to evaluate

(a) $6\mathbf{AB} - 9\mathbf{C}$ (b) $(\mathbf{AB})^3 + 5\mathbf{C}^2$
(c) $6\mathbf{AFB} - 2\mathbf{C}^3$ (d) $4\mathbf{BAF}$
(e) $(\mathbf{BCAF})^4$ (f) $2\mathbf{CA} + 37.3\mathbf{A}$
(g) $(\mathbf{CAB})^2$ (h) $0.73\mathbf{BA} + 1.6\mathbf{F}^6$

10.3 The Transpose Matrix

We continue the development of Section 10.2 by introducing the "transpose" of a matrix. Given any $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, we define the **transpose** of $\mathbf{A}$, denoted as $\mathbf{A}^{\mathrm{T}}$ and read as "$\mathbf{A}$-transpose," as

$$\mathbf{A}^{\mathrm{T}} \equiv \{a_{ji}\} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}, \tag{1}$$

that is, the $n \times m$ matrix is obtained by interchanging the rows and columns of $\mathbf{A}$: the first row of $\mathbf{A}$ becomes the first column of $\mathbf{A}^{\mathrm{T}}$, the second row of $\mathbf{A}$ becomes the second column of $\mathbf{A}^{\mathrm{T}}$, and so on. Or, the first column of $\mathbf{A}$ becomes the first row of $\mathbf{A}^{\mathrm{T}}$, and so on. That is, if we denote the i, j elements of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ as a_{ij} and a_{ij}^{T}, respectively, then

$$a_{ij}^{\mathrm{T}} = a_{ji}. \tag{2}$$

Be clear that $\mathbf{A}^{\mathrm{T}}$ is not $\mathbf{A}$ to the Tth power; it is the transpose of $\mathbf{A}$.

EXAMPLE 1. If

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 3 & 5 \\ 4 & 6 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 \\ 6 \\ 7 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = [1, -8, 9],$$

then

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 2 & -1 & 4 \\ 0 & 3 & 6 \\ 1 & 5 & 7 \end{bmatrix}, \quad \mathbf{B}^{\mathrm{T}} = [2, 6, 7], \quad \text{and} \quad \mathbf{C}^{\mathrm{T}} = \begin{bmatrix} 1 \\ -8 \\ 9 \end{bmatrix}. \quad ■$$

THEOREM 10.3.1 *Properties of the Transpose*

$$\left(\mathbf{A}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{A}, \tag{3a}$$

$$(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}, \tag{3b}$$

$$(\alpha \mathbf{A})^{\mathrm{T}} = \alpha \mathbf{A}^{\mathrm{T}}, \tag{3c}$$

$$(\mathbf{A}\mathbf{B})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \tag{3d}$$

where it is assumed in (3b) that $\mathbf{A}$ and $\mathbf{B}$ are conformable for addition, and in (3d) that they are conformable for multiplication.

Proof: Proof of (3a)–(3c) is left for the exercises. To prove (3d), let $\mathbf{AB} \equiv \mathbf{C} = \{c_{ij}\}$. By the definition of matrix multiplication,

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{4}$$

Thus,

$$c_{ij}^{\mathrm{T}} = c_{ji} = \sum_{k=1}^{n} a_{jk} b_{ki} = \sum_{k=1}^{n} b_{ki} a_{jk} = \sum_{k=1}^{n} b_{ik}^{\mathrm{T}} a_{kj}^{\mathrm{T}}. \tag{5}$$

Having returned, at the end of (5), to the pattern $(\,)_{ij} = \sum (\,)_{ik} (\,)_{kj}$ as in (4), we can conclude from (5) that

$$\mathbf{C}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \quad \text{or} \quad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}},$$

as was to be proved. Understand that the third equality in (5) is not equivalent to the matrix statement $\mathbf{AB} = \mathbf{BA}$ which, we recall from Section 10.2, is generally untrue. It is simply the scalar statement that $a_{jk} b_{ki} = b_{ki} a_{jk}$, which is true because the multiplication of scalars is commutative [e.g., $(2)(3) = (3)(2) = 6$]. ■

The striking feature of (3d) is the reversal in the order: $(\mathbf{AB})^{\mathrm{T}}$ on the left, $\mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}$ on the right. Notice how (3d) checks "dimensionally":

$$\begin{aligned} [(m \times n)(n \times p)]^{\mathrm{T}} &= (n \times p)^{\mathrm{T}} (m \times n)^{\mathrm{T}} \\ (m \times p)^{\mathrm{T}} &= (p \times n)(n \times m) \\ (p \times m) &= (p \times m). \end{aligned} \tag{6}$$

Naturally, (6) does not *prove* (3d), but it provides a useful check, just as we check the *physical* units (such as force, mass, length, and time) of an equation to be sure that they are consistent.

EXAMPLE 2. If $\mathbf{A} = \begin{bmatrix} 4 & 2 & -5 \\ 0 & 1 & 3 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 6 \\ 0 \\ -1 \end{bmatrix}$, say, then

$$\mathbf{AB} = \begin{bmatrix} 4 & 2 & -5 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 29 \\ -3 \end{bmatrix}, \quad \text{so} \quad (\mathbf{AB})^{\mathrm{T}} = [29, -3]$$

and

$$\mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} = [6, 0, -1] \begin{bmatrix} 4 & 0 \\ 2 & 1 \\ -5 & 3 \end{bmatrix} = [29, -3],$$

in agreement with (3d). ■

Furthermore, it follows from (3d) that

$$(\mathbf{ABC})^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}} \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}, \qquad (\mathbf{ABCD})^{\mathrm{T}} = \mathbf{D}^{\mathrm{T}} \mathbf{C}^{\mathrm{T}} \mathbf{B}^{\mathrm{T}} \mathbf{A}^{\mathrm{T}}, \tag{7}$$

and so on (Exercise 2).

Using lowercase boldface letters for matrices that happen to be column vectors from now on, let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Then the standard dot product

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{j=1}^{n} x_j y_j$$

can be expressed compactly, in matrix language, as*

$$\boxed{\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}} \mathbf{y}} \tag{8}$$

or, equivalently, as $\mathbf{y}^{\mathrm{T}}\mathbf{x}$, although not as $\mathbf{x}\mathbf{y}^{\mathrm{T}}$ or $\mathbf{y}\mathbf{x}^{\mathrm{T}}$, which expressions represent $n \times n$ matrices!

EXAMPLE 3. If $\mathbf{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$, say, then

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y} = [3, 1] \begin{bmatrix} 5 \\ 2 \end{bmatrix} = 15 + 2 = 17,$$

or

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{x} = [5, 2] \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 15 + 2 = 17,$$

whereas

$$\mathbf{x}\mathbf{y}^{\mathrm{T}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} [5, 2] = \begin{bmatrix} 15 & 6 \\ 5 & 2 \end{bmatrix} \neq \mathbf{x} \cdot \mathbf{y},$$

and

$$\mathbf{y}\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} 5 \\ 2 \end{bmatrix} [3, 1] = \begin{bmatrix} 15 & 5 \\ 6 & 2 \end{bmatrix} \neq \mathbf{x} \cdot \mathbf{y}. \quad ▮$$

Finally, two more definitions: if

$$\mathbf{A}^{\mathrm{T}} = \mathbf{A} \tag{9}$$

we say that $\mathbf{A}$ is **symmetric**, and if

$$\mathbf{A}^{\mathrm{T}} = -\mathbf{A}, \tag{10}$$

*There is a small difficulty here: $\mathbf{x} \cdot \mathbf{y}$ is to be scalar, whereas $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ is a 1×1 matrix. Thus, what we really intend, by $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ in (8), is not the 1×1 matrix, but rather the scalar element inside it. That could be noted by writing $\mathbf{x} \cdot \mathbf{y} = (\mathbf{x}^{\mathrm{T}}\mathbf{y})_{11}$ instead, but it will be simpler to leave (8) intact, with the *understanding* that the right-hand side is a scalar.

we say that it is **skew-symmetric** (or antisymmetric). For either of these properties to apply $\mathbf{A}$ must be square, since otherwise $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}$ would be of different form. And for $\mathbf{A}$ to be skew-symmetric all of its diagonal elements must be zero, since (10) implies that $a_{ji} = -a_{ij}$, or if we set $i = j$, $a_{ii} = -a_{ii}$; thus $2a_{ii} = 0$, so $a_{ii} = 0$ for each i.

It would be reasonable to imagine that the likelihood of encountering purely symmetric or skew-symmetric matrices in applications would be slim. On the contrary, we shall see that symmetric matrices arise frequently, and that their symmetry is often a consequence of fundamental physical principles, rather than chance.

Closure. The key points in this section are the defining of the transpose of any matrix, and the results $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{x}\cdot\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ (or $\mathbf{y}^{\mathrm{T}}\mathbf{x}$). Note that the transpose notation is sometimes used to save space. For instance, it takes less vertical space on the page to write $\mathbf{x}^{\mathrm{T}} = [7, 2]$ than $\mathbf{x} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$.

Computer software. The relevant *Maple* function, to take the transpose of a matrix $\mathbf{A}$, is the command **transpose**, within the linalg package. For instance, to obtain the transpose of $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ enter

```
with(linalg):
```

to access the transpose(A) command. Enter

```
A := array([[1, 2, 3], [4, 5, 6]]);
```

and return, then enter

```
transpose(A);
```

and return. The output is

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

EXERCISES 10.3

1. (a) If $\mathbf{x} = [3, -3]^{\mathrm{T}}$ and $\mathbf{y} = [1, 2]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.
(b) If $\mathbf{x} = [4, -1, 0]^{\mathrm{T}}$ and $\mathbf{y} = [1, 2, 3]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.
(c) If $\mathbf{x} = [0, 4, -2, 1]^{\mathrm{T}}$ and $\mathbf{y} = [3, 0, 1, -2]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.

2. Show that (7) follows from (3d).

3. Recall that in general $\mathbf{AB} \neq \mathbf{BA}$, and that a *necessary* (but not sufficient) condition for equality to hold is that both $\mathbf{A}$ and $\mathbf{B}$ be square and of the same order. Perhaps a *sufficient* condition is that $\mathbf{A}$ and $\mathbf{B}$ both be of the same order and *symmetric*. Prove or disprove this hypothesis.

4. Verify $(\mathbf{ABC})^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ directly, for

(a) $\mathbf{A} = \begin{bmatrix} 5 & -2 \\ 0 & 1 \\ 1 & 3 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 \\ 7 \end{bmatrix}$, $\mathbf{C} = [3, 1, 2, 9]$

(b) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 4 & -1 & 5 \\ 0 & 3 & 0 \end{bmatrix}$

(c) $\mathbf{A} = [5, 3, 0], \mathbf{B} = \begin{bmatrix} -2 & 1 \\ 6 & 4 \\ 5 & 3 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$

(d) $\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 3 & 4 \\ 1 & 2 \\ 0 & 1 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$

5. Prove the properties (3a), (3b), and (3c).

6. Even if a (square) matrix is neither symmetric nor skew-symmetric it can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix. Specifically, writing

$$\begin{aligned} \mathbf{A} &= \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{T}}) + \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathrm{T}}) \\ &\equiv \quad \mathbf{A}_1 \quad + \quad \mathbf{A}_2, \end{aligned} \tag{6.1}$$

show that $\mathbf{A}_1$ is symmetric and $\mathbf{A}_2$ is skew-symmetric. NOTE: Equation (6.1) is but one such decomposition. For instance, we saw in Chapter 9 that a vector $\mathbf{v}$ in 3-space can be decomposed into the sum of two vectors, one lying in a given plane (the "projection of $\mathbf{v}$" onto that plane) and the other perpendicular to that plane; when we study vector fields we will see that any vector field can be decomposed as the sum of two fields, one irrotational and the other solenoidal; when we study Fourier series we will see that any function $f(x)$ can be decomposed as the sum of two functions, one even and the other odd; and so on. Thus, such decompositions are not uncommon in mathematics.

7. Decompose the given matrix as the sum of two matrices, one symmetric and one skew symmetric, as explained in Exercise 6.

(a) $\begin{bmatrix} 3 & 2 \\ 1 & -5 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 8 & -2 \\ 4 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 9 & 8 & 7 \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 3 & 0 & 6 \end{bmatrix}$

8. (*Quadratic forms*) The quadratic function $ax^2 + by^2 + cxy$ is said to be a **quadratic form** in x and y, $ax^2 + by^2 + cz^2 + dxy + exz + fyz$ is a quadratic form in x, y, z, and so on. Every quadratic form in the n variables $x_1, \dots, x_n$ can be expressed, in matrix notation, as $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric $n \times n$ matrix. For each given quadratic form, determine the $\mathbf{A}$ matrix. NOTE: Quadratic forms will be important to us in Chapter 11.

(a) $6x_1^2 + x_2^2 - 8x_1x_2$ in x_1, x_2

(b) $x_1^2 - 3x_2^2 + 6x_1x_2$ in x_1, x_2

(c) $4x_1^2 + x_2^2 - x_3^2 + 8x_1x_2 + 3x_1x_3 - 2x_2x_3$ in x_1, x_2, x_3

(d) $x_1^2 - 4x_3^2 + 2x_1x_3 - 10x_2x_3$ in x_1, x_2, x_3

(e) $3x_2^2 + x_3^2 - 6x_1x_3$ in x_1, x_2, x_3

(f) $x_1^2 + x_2^2 + x_4^2 + x_1x_3 + x_1x_4 + x_2x_3$ in x_1, x_2, x_3, x_4

9. Show that if $\mathbf{A}$ is an $m \times n$ matrix, then $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is symmetric. HINT: There is often an inclination to work out a problem like this using "brute force," i.e., by actually writing out the $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ matrices, multiplying them, and examining the resulting matrix to see if it is symmetric. Whenever possible, we advise against such an approach. In this problem, for example, we wish to show that $\mathbf{C}^{\mathrm{T}} = \mathbf{C}$, where $\mathbf{C}$ is short for $\mathbf{A}\mathbf{A}^{\mathrm{T}}$; i.e., we wish to show that $(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$, and this can be done (in one or two short lines) using the properties stated in Theorem 10.3.1.

10. Prove that *the product* $\mathbf{AB}$ *need not be symmetric, even if* $\mathbf{A}$ *and* $\mathbf{B}$ *are both symmetric and of the same order.*

11. Let

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 0 & 5 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -4 & 2 \\ 8 & 1 & 4 \end{bmatrix}.$$

Use computer software, such as *Maple*, to evaluate

(a) $\mathbf{A}\mathbf{B}^{\mathrm{T}}$ (b) $\mathbf{B}\mathbf{A}^{\mathrm{T}}$

(c) $(\mathbf{A}\mathbf{B}^{\mathrm{T}})^5$ (d) $(\mathbf{B}\mathbf{A}^{\mathrm{T}})^{\mathrm{T}}$

(e) $(\mathbf{B}^{\mathrm{T}}\mathbf{A})^8$ (f) $2\mathbf{A}^{\mathrm{T}} - 7.3\mathbf{B}^{\mathrm{T}}$

10.4 Determinants

In this section we introduce a scalar quantity associated with every square matrix, the so-called "determinant" of the matrix. We denote the **determinant** of an $n \times n$

matrix $\mathbf{A} = \{a_{ij}\}$ as

$$\det\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}, \tag{1}$$

that is, with straight line braces instead of square brackets. Determinants are prominent in Chapter 3 in connection with the linear dependence or independence of sets of functions, especially with sets of solutions of linear homogeneous differential equations. More generally, they play a key role in the theory of systems of linear algebraic equations, as discussed in Section 10.5.

The determinant of an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is defined by the **cofactor expansion**

$$\boxed{\det\mathbf{A} \equiv \sum_{1}^{n} a_{jk}A_{jk},} \tag{2}$$

where *the summation is carried out on j for any fixed value of k* $(1 \leq k \leq n)$ *or on k for any fixed value of j* $(1 \leq j \leq n)$. A_{jk} is called the **cofactor** of the a_{jk} element and is defined as

$$A_{jk} \equiv (-1)^{j+k}M_{jk}, \tag{3}$$

where M_{jk} is called the **minor** of a_{jk}, namely, the determinant of the $(n-1) \times (n-1)$ matrix that survives when the row and column containing a_{jk} (the jth row and the kth column) are struck out.

For example, if

$$\mathbf{A} = \begin{bmatrix} 4 & 7 & -2 \\ 0 & 3 & 2 \\ 1 & 5 & 6 \end{bmatrix},$$

then

$$M_{11} = \begin{vmatrix} 3 & 2 \\ 5 & 6 \end{vmatrix}, \quad M_{12} = \begin{vmatrix} 0 & 2 \\ 1 & 6 \end{vmatrix}, \quad \text{and} \quad M_{23} = \begin{vmatrix} 4 & 7 \\ 1 & 5 \end{vmatrix}.$$

Thus, if $\mathbf{A}$ is $n \times n$, then the right-hand side of (2) is a linear combination of n determinants, each of which is $(n-1) \times (n-1)$. Each of these, in turn, may be expressed as a linear combination of $(n-2) \times (n-2)$ determinants, and so on, until we have a (perhaps large) number of 1×1 determinants. Thus, the definition (2) is logically incomplete until we define a 1×1 determinant, which we do as follows:

$$\det\left[\ a_{11}\ \right] = \left|\ a_{11}\ \right| \equiv a_{11}. \tag{4}$$

That is, the determinant of a 1×1 matrix $\left[\ a_{11}\ \right]$ is simply a_{11} itself. CAUTION: In the present context, the braces around a_{11}, in the middle member of (4), denote determinant, not absolute value. For instance, $\det[-6] = |-6| = -6$.

Using (2) and (4), let us work out the determinant of any 2×2 matrix. Recalling that we can sum on j with k fixed, or vice versa, let us sum on k with $j = 1$, say. Then

$$\begin{aligned} \det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} &= \sum_{k=1}^{2} a_{1k} A_{1k} \\ &= a_{11}A_{11} + a_{12}A_{12} = a_{11}(-1)^{1+1}M_{11} + a_{12}(-1)^{1+2}M_{12} \\ &= a_{11}(+1)\,|a_{22}| + a_{12}(-1)\,|a_{21}| \\ &= a_{11}a_{22} - a_{12}a_{21}, \end{aligned} \tag{5}$$

which result is probably familiar to you from earlier studies. Observe that the $(-1)^{j+k}$ in (3) is simply $+1$ if $j + k$ is even, and -1 if $j + k$ is odd.

EXAMPLE 1. Let

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & -1 \\ 4 & 3 & 5 \\ 2 & 0 & -4 \end{bmatrix}. \tag{6}$$

Using (2) with $j = 1$, say,

$$\begin{aligned} \det\mathbf{A} = \sum_{k=1}^{3} a_{1k} A_{1k} &= a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13} \\ &= a_{11}(+1)M_{11} + a_{12}(-1)M_{12} + a_{13}(+1)M_{13} \\ &= a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} \\ &= (0)\begin{vmatrix} 3 & 5 \\ 0 & -4 \end{vmatrix} - (2)\begin{vmatrix} 4 & 5 \\ 2 & -4 \end{vmatrix} + (-1)\begin{vmatrix} 4 & 3 \\ 2 & 0 \end{vmatrix} \\ &= 0 - (2)(-16 - 10) + (-1)(0 - 6) = 58. \end{aligned} \tag{7}$$

This particular choice ($j = 1$, summing on k) is said to be *cofactor expansion about the first row* since (7) is the sum of the first-row elements a_{11}, a_{12}, a_{13}, multiplied by their cofactors.

According to (2), we can expand about *any* row or column. Let us illustrate that the same answer is indeed obtained if we expand about other rows or columns.

Expansion about second row [i.e., set $j = 2$ in (2), and sum on k]:

$$\begin{aligned} \det\mathbf{A} = \sum_{k=1}^{3} a_{2k} A_{2k} &= a_{21}A_{21} + a_{22}A_{22} + a_{23}A_{23} \\ &= a_{21}(-1)M_{21} + a_{22}(+1)M_{22} + a_{23}(-1)M_{23} \\ &= -(4)\begin{vmatrix} 2 & -1 \\ 0 & -4 \end{vmatrix} + (3)\begin{vmatrix} 0 & -1 \\ 2 & -4 \end{vmatrix} - (5)\begin{vmatrix} 0 & 2 \\ 2 & 0 \end{vmatrix} \\ &= -(4)(-8) + (3)(2) - (5)(-4) = 58, \end{aligned}$$

again.

Expansion about third column ($k = 3$, sum on j):

$$\begin{aligned}\det\mathbf{A} = \sum_{j=1}^{3} a_{j3}A_{j3} &= a_{13}A_{13} + a_{23}A_{23} + a_{33}A_{33} \\ &= a_{13}(+1)M_{13} + a_{23}(-1)M_{23} + a_{33}(+1)M_{33} \\ &= (-1)\begin{vmatrix} 4 & 3 \\ 2 & 0 \end{vmatrix} - (5)\begin{vmatrix} 0 & 2 \\ 2 & 0 \end{vmatrix} + (-4)\begin{vmatrix} 0 & 2 \\ 4 & 3 \end{vmatrix} \\ &= (-1)(-6) - (5)(-4) + (-4)(-8) = 58,\end{aligned}$$

once more. ▮

Since we may expand about any row or column, it is convenient in hand calculations to choose the row or column with the most zeros in it since those terms in the expansion then drop out.

Notice carefully that for large n the cofactor expansion process is exceedingly laborious. Even if $n = 10$, say, which is still quite modest, (2) gives a linear combination of ten 9×9 determinants. In turn, each of these ten is evaluated as a linear combination of nine 8×8 determinants, and, so on! Let us see just how serious this predicament is. For estimating purposes, let us count each multiplication, addition, and subtraction as one "calculation." It can be shown (Exercise 18a) that the number of calculations $N(n)$ required in the evaluation (by cofactor expansion) of an $n \times n$ determinant is

$$N(n) \sim e\,n! \tag{8}$$

as $n \to \infty$, where $e \approx 2.718$ is the base of the natural logarithm, and $n!$ is n factorial. If each calculation takes approximately one microsecond, then some time estimates are as follows. (Before reading on, we urge you to guess how long such a computer would take to evaluate a 25×25 determinant.)

n	Computing Time
5	0.0003 sec
10	10 sec
15	4×10^6 sec ≈ 40 days
20	7×10^{12} sec $\approx 210{,}000$ years
25	4×10^{19} sec $\approx 10^{12}$ years

It is interesting that faster computers offer no hope. For instance, even a computer that is a million times as fast would still take around 10^6 years to evaluate a 25×25 determinant. And scientific calculations can easily involve determinants that are 250×250.

It is tempting to conclude that "determinants are worthless," but let us see if we can come up with a more efficient algorithm than the cofactor expansion. A logical starting point is to first determine the various properties of determinants so that we can use them to design a better algorithm.

First, we need to introduce the idea of a "triangular" matrix. A square matrix $\mathbf{A} = \{a_{ij}\}$ is **upper triangular** if $a_{ij} = 0$ for all $j < i$ and **lower triangular** if $a_{ij} = 0$ for all $j > i$. That is,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & a_{nn} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \vdots \\ a_{n1} & & \cdots & a_{nn} \end{bmatrix} \tag{9}$$

are upper triangular and lower triangular, respectively. If a matrix is upper triangular or lower triangular it is said to be **triangular**.

Here are the properties that we will need.

PROPERTIES OF DETERMINANTS

D1. If any row (or column) of $\mathbf{A}$ is modified by adding α times the corresponding elements of another row (or column) to it, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = \det\mathbf{A}$.
Symbolically: $\mathbf{r}_j \to \mathbf{r}_j + \alpha\mathbf{r}_k$

D2. If any two rows (or columns) of $\mathbf{A}$ are interchanged, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = -\det\mathbf{A}$.
Symbolically: $\mathbf{r}_j \leftrightarrow \mathbf{r}_k$

D3. If $\mathbf{A}$ is triangular, then $\det\mathbf{A}$ is simply the product of the diagonal elements, $\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}$.

Of these, D3 is easily proved. For consider the general upper triangular matrix in (9). Doing a cofactor expansion about the first column gives a_{11} times an $(n-1)\times(n-1)$ minor determinant, which is again upper triangular. Expanding the latter about its first column gives a_{22} times an $(n-2)\times(n-2)$ minor determinant, which is again upper triangular. Repeating the process leads to $\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}$. Similarly for the general lower triangular matrix in (9), except that in that case we expand about the first row, repeatedly, rather than the first column.

Let us illustrate the use of the properties D1–D3 instead of the cofactor expansion.

EXAMPLE 2. Consider the $\mathbf{A}$ matrix of Example 1 again.

$$\det\mathbf{A} = \begin{vmatrix} 0 & 2 & -1 \\ 4 & 3 & 5 \\ 2 & 0 & -4 \end{vmatrix} = -\begin{vmatrix} 2 & 0 & -4 \\ 4 & 3 & 5 \\ 0 & 2 & -1 \end{vmatrix}$$

$$= -\begin{vmatrix} 2 & 0 & -4 \\ 0 & 3 & 13 \\ 0 & 2 & -1 \end{vmatrix} = -\begin{vmatrix} 2 & 0 & -4 \\ 0 & 3 & 13 \\ 0 & 0 & -\frac{29}{3} \end{vmatrix} = -(2)(3)\left(-\frac{29}{3}\right) = 58,$$

as obtained in Example 1. In the second equality we interchanged the first and third rows ($\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$), thereby changing the sign of the determinant (D2) so we compensated by putting the minus sign out in front. In the third equality we modified the second row by adding -2 times the first row to it ($\mathbf{r}_2 \to \mathbf{r}_2 - 2\mathbf{r}_1$), which step left the determinant unchanged (D1). In the fourth equality we modified the third row by adding $-\frac{2}{3}$ times the second row to it ($\mathbf{r}_3 \to \mathbf{r}_3 - \frac{2}{3}\mathbf{r}_2$), which step left the determinant unchanged (D1). Since those steps produced a triangular matrix, we could then use D3. ▮

The point, then, is to use some combination of D1 and D2 steps to reduce the determinant to triangular form, in which case it is evaluated easily by D3. Of course the method is quite similar to Gauss elimination, described in Section 8.3. For instance, compare D1 and D2 with the first and third elementary equation operations listed in Section 8.3. For reference purposes, we will call the method illustrated in Example 2 the method of **triangularization**.

It is hard to tell, from the 3×3 calculation in Example 2, whether the method is more efficient than the cofactor expansion. However, in Exercise 18b it is shown that using triangularization the number of calculations $N(n)$ is

$$N(n) \sim \frac{2n^3}{3} \tag{10}$$

as $n \to \infty$. Again assuming one microsecond per calculation, (10) gives a computing time of around 0.005 second for $n = 20$ and 0.01 second for $n = 25$, compared with 210,000 years and 10^{12} years, respectively! [Comparing (8) and (10), we can see how much faster $n!$ grows than n^3.]

The upshot is that except for small hand calculations we should avoid the cofactor expansion, and should use triangularization instead.

Although properties D1 – D3 suffice for the efficient calculation of determinants, other properties are sometimes useful as well, and are listed below.

ADDITIONAL PROPERTIES OF DETERMINANTS

D4. If all the elements of any row or column are zero, then $\det\mathbf{A} = 0$.

D5. If any two rows or columns are proportional to each other, then $\det\mathbf{A} = 0$.

D6. If any row (column) is a linear combination of other rows (columns), then $\det\mathbf{A} = 0$.

D7. If all the elements of any row or column are scaled by α, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = \alpha\det\mathbf{A}$.

D8. $\det(\alpha\mathbf{A}) = \alpha^n\det\mathbf{A}$.

D9. If any one row (or column) $\mathbf{a}$ of $\mathbf{A}$ is separated as $\mathbf{a} = \mathbf{b} + \mathbf{c}$, then

$$\det\mathbf{A}|_\mathbf{a} = \det\mathbf{A}|_\mathbf{b} + \det\mathbf{A}|_\mathbf{c},$$

where $\mathbf{A}|_\mathbf{a}$ denotes the $\mathbf{A}$ matrix with $\mathbf{a}$ intact, $\mathbf{A}|_\mathbf{b}$ denotes the $\mathbf{A}$ matrix with $\mathbf{b}$ in place of $\mathbf{a}$, and similarly for $\mathbf{A}|_\mathbf{c}$. For example,

$$\begin{vmatrix} 6+2 & -3+1 & 5+4 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix} = \begin{vmatrix} 6 & -3 & 5 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix} + \begin{vmatrix} 2 & 1 & 4 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix}.$$

D10. The determinant of $\mathbf{A}$ and its transpose are equal,

$$\det(\mathbf{A}^{\mathrm{T}}) = \det\mathbf{A}.$$

D11. In general,

$$\det(\mathbf{A}+\mathbf{B}) \neq \det\mathbf{A} + \det\mathbf{B}.$$

D12. The determinant of a product equals the product of their determinants,

$$\boxed{\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B}).} \tag{11}$$

These properties are not independent of each other. For example, D5 follows from D1 and D4, and D4 follows from D6. Keep in mind that det( ) is *not linear*. That is, if α, β are scalars and $\mathbf{A}$, $\mathbf{B}$ are $n \times n$ matrices, then

$$\boxed{\det(\alpha\mathbf{A} + \beta\mathbf{B}) \neq \alpha\det\mathbf{A} + \beta\det\mathbf{B},} \tag{12}$$

in general. For instance, if $\beta = 0$ is $\det(\alpha\mathbf{A}) = \alpha\det\mathbf{A}$? No, according to D7 it is $\alpha^n\det\mathbf{A}$. Or, with $\alpha = \beta = 1$, is $\det(\mathbf{A}+\mathbf{B}) = \det\mathbf{A} + \det\mathbf{B}$? Not in general, according to D11. This result may come as a surprise since we are studying "linear algebra." Also surprising is the truth of D12, if we contrast the complexity of the matrix multiplication $\mathbf{AB}$ on the left with the simplicity of the outcome, expressed on the right-hand side. This result was proved by Cauchy in 1815.*

Closure. Every $n \times n$ matrix $\mathbf{A}$ has a value associated with it called its determinant and denoted as $\det\mathbf{A}$. Though $\det\mathbf{A}$ is defined, traditionally, by the cofactor expansion (2), we find that that formula is useless, computationally, unless n is quite small. Thus, we study various properties of the determinant and put forward a computational algorithm called triangularization, based upon properties D1–D3, that is incredibly efficient compared to the cofactor expansion.

Computer software. Using *Maple*, one can evaluate determinants using the **det(A)** command. For instance, to evaluate the determinant of the matrix $\mathbf{A}$ given by (6), enter

```
with(linalg):
```

to access the det(A) command. Then enter

```
det([[0, 2, −1], [4, 3, 5], [2, 0, −4]]);
```

and return. The output is 58. Alternatively, the sequence

```
with(linalg):
A := array([[0, 2, −1], [4, 3, 5], [2, 0, −4]]) :
det(A);
```

***Augustin–Louis Cauchy* (1789–1857) is among the great mathematicians. Unlike his contemporary, Gauss who published little of his work, Cauchy published more than 700 papers. Among the subjects on which he worked were determinants, ordinary and partial differential equations, complex variable theory, and the wave theory of light.

gives the same result.

EXERCISES 10.4

1. In (5) we evaluated the determinant of a general 2×2 matrix using a cofactor expansion about the first row. Evaluate it again, using a cofactor expansion about the second row instead, then about the first column, and then about the second column, showing that the answer is the same in each case.

2. Evaluate each, using a cofactor expansion about the first and last rows, and also about the last column.

(a) $\begin{vmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 1 & 1 & 1 \end{vmatrix}$ (b) $\begin{vmatrix} 2 & -3 & 0 \\ 1 & 4 & 2 \\ -6 & 1 & 5 \end{vmatrix}$

(c) $\begin{vmatrix} -4 & 1 & 0 \\ 3 & 2 & 0 \\ 1 & 5 & 7 \end{vmatrix}$ (d) $\begin{vmatrix} 3 & 3 & 12 \\ 0 & 6 & -1 \\ 4 & 0 & 0 \end{vmatrix}$

(e) $\begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{vmatrix}$ (f) $\begin{vmatrix} -5 & 2 & 1 & 0 \\ 4 & 0 & 3 & 0 \\ 2 & 4 & 7 & 1 \\ 0 & 0 & -2 & 6 \end{vmatrix}$

(g) $\begin{vmatrix} 2 & 0 & 1 & 0 \\ 0 & 3 & 1 & -1 \\ 0 & 4 & 5 & 0 \\ 1 & 2 & 3 & 6 \end{vmatrix}$ (h) $\begin{vmatrix} 0 & 1 & 2 & 0 \\ 3 & -1 & 1 & 4 \\ 5 & 6 & -7 & 1 \\ 0 & 2 & 1 & 0 \end{vmatrix}$

(i) $\begin{vmatrix} a & 0 & 0 & 0 \\ 0 & b & c & 0 \\ 0 & d & e & 0 \\ 0 & 0 & 0 & f \end{vmatrix}$ (j) $\begin{vmatrix} a & b & c & 0 \\ d & e & f & 0 \\ g & h & i & 0 \\ 0 & 0 & 0 & k \end{vmatrix}$

3. (a)–(j) Same as Exercise 2, but expanding about the second row, and about the first column.

4. (a)–(h) Same as Exercise 2, but using the method of triangularization.

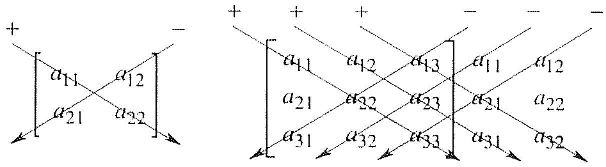
5. (a)–(h) Same as Exercise 2, but using computer software.

6. Evaluate, by any means other than computer software, showing your steps or logic. You may use any of the properties D1–D12.

(a) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 0 & 1 & -3 & 5 \end{vmatrix}$ (b) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 0 \end{vmatrix}$

(c) $\begin{vmatrix} 0 & 0 & a \\ 0 & b & c \\ d & e & f \end{vmatrix}$ (d) $\begin{vmatrix} a & b & c \\ d & e & 0 \\ f & 0 & 0 \end{vmatrix}$

7. A mnemonic device often put forward for evaluating 2×2 and 3×3 determinants is as shown below.



In other words, the determinants are the sums of the indicated products, with each product carrying the indicated sign. For example, in the 2×2 case this device gives

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = +(a_{11}a_{22}) - (a_{21}a_{12}),$$

which does agree with (5). We now state the problem: write out the mnemonic result for the 3×3 case, and verify (by cofactor expansion) that it is correct. CAUTION: This device does *not* hold, in general, for $n \times n$ determinants if $n \geq 4$.

8. Let an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ be diagonal. Show that

$$\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}. \tag{8.1}$$

9. (a) Suppose that an $n \times n$ matrix $\mathbf{A}$ can be partitioned into the **block-diagonal** form

$$\mathbf{A} = \begin{bmatrix} \boxed{\mathbf{A}_1} & 0 & \cdots & 0 \\ 0 & \boxed{\mathbf{A}_2} & \cdots & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \boxed{\mathbf{A}_m} \end{bmatrix},$$

where $\mathbf{A}_1, \dots, \mathbf{A}_m$ are all square, although not necessarily all of the same order. Show that

$$\det\mathbf{A} = (\det\mathbf{A}_1)(\det\mathbf{A}_2)\cdots(\det\mathbf{A}_m). \tag{9.1}$$

This result may be regarded as a generalization of (8.1), above, wherein $\mathbf{A}_1, \dots, \mathbf{A}_m$ were all 1×1's.

(b) Does (9.1) still hold if the elements above the m blocks are

nonzero? Explain.
(c) Does (9.1) still hold if the elements below the m blocks are nonzero? Explain.
(d) To which determinants, in Exercise 2, can (9.1) be applied? In each of those cases use (9.1) to evaluate $\det\mathbf{A}$.

10. Deduce, from property D12, that if $\mathbf{A}_1, \dots, \mathbf{A}_k$ are $n \times n$ matrices, then

$$\det(\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_k) = (\det\mathbf{A}_1)(\det\mathbf{A}_2)\cdots(\det\mathbf{A}_k). \quad (10.1)$$

11. (a) Derive the property D7. HINT: Write out the cofactor expansion about the row or column in question.
(b) Then, show that D8 follows from D7.

12. Prove property D6, using any of the other listed properties.

13. Prove property D9. HINT: Write out the cofactor expansion about the row (or column) $\mathbf{a}$.

14. (*Routh–Hurwitz criterion*) First, review Section 3.4.5, on stability. According to the **Routh–Hurwitz criterion**, necessary and sufficient conditions for the stability of the system governed by equation (65) in Section 3.4.5 (i.e., for all the roots of its characteristic equation to have negative real parts) are that $a_j > 0$ for each $j = 1, \dots, n$, and that $\Delta_j > 0$ for each $j = 1, \dots, n$, where

$$\Delta_j = \begin{vmatrix} a_1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ a_{2j-1} & a_{2j-2} & a_{2j-3} & a_{2j-4} & & \cdots & a_j \end{vmatrix}.$$

Zeros are entered for any a_k's that are called for where $k > n$. For example, if $n = 3$, then

$$\Delta_1 = a_1, \quad \Delta_2 = \begin{vmatrix} a_1 & 1 \\ a_3 & a_2 \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} a_1 & 1 & 0 \\ a_3 & a_2 & a_1 \\ 0 & 0 & a_3 \end{vmatrix};$$

expanding Δ_3 about the third row yields the simplification $\Delta_3 = a_3\Delta_2$. Here is the problem: Use the Routh–Hurwitz criterion to determine whether or not the systems associated with the following characteristic equations are stable.

(a) $\lambda^4 + 6\lambda^3 + 5\lambda^2 + 4\lambda + 1 = 0$
(b) $\lambda^4 + 2\lambda^3 + 7\lambda^2 + 4\lambda + 8 = 0$
(c) $\lambda^4 + 2\lambda^3 + 5\lambda^2 + 8\lambda + 12 = 0$
(d) $\lambda^4 + \lambda^3 + 4\lambda + 8 = 0$
(e) $\lambda^5 + \lambda^4 + \lambda^3 + \lambda^2 + \lambda + 1 = 0$
(f) $\lambda^5 + \lambda^4 + \lambda^3 + \lambda^2 + \lambda + 8 = 0$
(g) $\lambda^5 + 2\lambda^4 + 3\lambda^3 + 4\lambda^2 + 5\lambda + 6 = 0$

15. It can be shown that the equations

$$\begin{aligned} a_0x^2 + a_1x + a_2 &= 0, \qquad (a_0 \neq 0) \\ b_0x^2 + b_1x + b_2 &= 0 \qquad (b_0 \neq 0) \end{aligned}$$

have a common root if and only if

$$\begin{vmatrix} a_0 & a_1 & a_2 & 0 \\ 0 & a_0 & a_1 & a_2 \\ b_0 & b_1 & b_2 & 0 \\ 0 & b_0 & b_1 & b_2 \end{vmatrix} = 0.$$

[Similar results for two algebraic equations of degrees m and n, say, were put forward by *James Joseph Sylvester* (1814–1897).] Use this result to determine whether or not the following equation pairs have any common roots.

(a) $3x^2 + 2x - 5 = 0$ (b) $3x^2 + 2x - 5 = 0$
$3x^2 + 3x - 2 = 0$ $x^2 + x + 1 = 0$

16. (a) Suppose that the elements a_{ij} of an $n \times n$ matrix $\mathbf{A}$ are differentiable functions of some parameter t. Regarding $\det\mathbf{A}$ as a function of the n^2 variables $a_{11}, a_{12}, \dots, a_{nn}$, show that

$$\frac{\partial}{\partial a_{ij}}(\det\mathbf{A}) = A_{ij}, \qquad (16.1)$$

where A_{ij} is the cofactor of a_{ij}. Then use (16.1) and chain differentiation to show that

$$\frac{d}{dt}(\det\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\frac{da_{ij}}{dt}, \qquad (16.2)$$

a formula first given by *Carl Gustav Jacob Jacobi* (1804–1851) in 1841. By the $\sum\sum$ notation we mean

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} \equiv \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right).$$

For example, if $n = 2$, then

$$\begin{aligned} \sum_{i=1}^{2}\sum_{j=1}^{2} c_{ij} = \sum_{i=1}^{2}\left(\sum_{j=1}^{2} c_{ij}\right) &= \sum_{i=1}^{2}(c_{i1} + c_{i2}) \\ &= c_{11} + c_{21} + c_{12} + c_{22}. \end{aligned}$$

Observe that (16.2) is equivalent to the statement

$$\frac{d}{dt}(\det\mathbf{A}) = \begin{vmatrix} \dfrac{da_{11}}{dt} & \cdots & \dfrac{da_{1n}}{dt} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}$$

$$+\cdots+\begin{vmatrix} a_{11} & \cdots & a_{1n}\\ \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n} \\ \dfrac{da_{n1}}{dt} & \cdots & \dfrac{da_{nn}}{dt}\end{vmatrix}. \qquad (16.3)$$

(b) Thus, evaluate $\dfrac{d}{dt}(\det\mathbf{A})$ if

$$\mathbf{A}=\begin{bmatrix} t^2 & t & 2 \\ 0 & 3t & 1 \\ 4 & 0 & \sin t\end{bmatrix},$$

and check your result by working out det**A** and taking its t derivative.

17. (*Vandermonde determinant*) First, review Theorem 3.4.1 and its proof. The determinant in (43),

$$\begin{vmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1}\end{vmatrix}, \qquad (17.1)$$

is known as a **Vandermonde determinant**. It can be shown that it equals $(-1)^{n(n-1)/2}\Pi$, where Π denotes the product of all factors $\lambda_j-\lambda_k$ with $j<k$ $(\leq n)$. For example, if $n=3$ then $\Pi=(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)(\lambda_2-\lambda_3)$. The key property of any $n\times n$ Vandermonde determinant, which can be seen from this result, is that it is nonzero if and only if all of the λ_j's are distinct. The problem that we pose is for you to verify that the determinant is equal to $(-1)^{n(n-1)/2}\Pi$, as claimed above, simply by working out the determinant, for the cases where $n=2$ and 3.

18. (a) Derive (8). HINT: Show that $N(n)=nN(n-1)+2n-1$ for $n\geq 2$. Let us use the subscript notation $N(n)=P_n$, which is generally used for functions of a discrete integer variable. Thus,

$$P_n-nP_{n-1}=2n-1, \qquad (n\geq 2) \qquad (18.1)$$

with the initial condition $P_2=3$ (two multiplications and one subtraction). Seeking P_n in the form $n!Q_n$, show that Q_n satisfies the difference equation

$$Q_n-Q_{n-1}=\frac{2n-1}{n!} \qquad (18.2)$$

with initial condition $Q_2=3/2$, which admits the solution

$$Q_n=\frac{3}{2}+2\sum_{j=2}^{n-1}\frac{1}{j!}-\sum_{j=3}^{n}\frac{1}{j!}. \qquad (18.3)$$

[You need merely verify (18.3).] Finally, show from (18.3) that $Q_n\sim e$ as $n\to\infty$.

(b) Derive (10). HINT: $n(n-1)+(n-1)(n-2)+\cdots+(2)(1)=n(n^2-1)/3$.

10.5 Rank; Application to Linear Dependence and to Existence and Uniqueness for Ax=c

With determinants defined, we can now introduce one more concept, the "rank" of a matrix, which concept will enable us to obtain important results regarding linear dependence, and also regarding the existence and uniqueness of solutions of the linear equation $\mathbf{Ax}=\mathbf{c}$.

10.5.1. Rank. First, we say that any matrix obtained from a given $m\times n$ matrix **A** by deleting at most $m-1$ rows and at most $n-1$ columns from **A**

is a **submatrix** of **A**. For instance, the 2 × 3 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

has 21 submatrices: one 2 × 3 (**A** itself), three 2 × 2's, two 1 × 3's, three 2 × 1's, six 1 × 2's, and six 1 × 1's. For instance,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix}, \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}, [a_{21}, a_{23}],$$

and $[a_{13}]$ are all submatrices of **A**.

Then the rank of a matrix is defined as follows.

DEFINITION 10.5.1 *Rank*
A matrix **A**, not necessarily square, is of **rank** r, or $r(\mathbf{A})$, if it contains at least one $r \times r$ submatrix with nonzero determinant but no square submatrix larger than $r \times r$ with nonzero determinant. A matrix is of rank 0 if it is a zero matrix.

EXAMPLE 1. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 & 0 \\ 0 & 3 & 3 & 6 \\ 1 & 4 & 5 & 9 \end{bmatrix}. \tag{1}$$

Certainly, r is at most 3 in this case since the largest possible square submatrix of **A** is 3 × 3. (More generally, if **A** is $m \times n$, then r is at most equal to the smaller of m and n.) However, upon calculation, we find that all four of the 3 × 3 submatrices have zero determinant so that r is at most 2. In fact, there are a number of 2 × 2 submatrices with nonzero determinant such as

$$\begin{vmatrix} 2 & -1 \\ 0 & 3 \end{vmatrix} = 6 \neq 0,$$

but even if there were only one such submatrix that would still be all we need to conclude that $r(\mathbf{A}) = 2$. ▮

EXAMPLE 2. The ranks of

$$\mathbf{A} = \begin{bmatrix} 5 \\ 6 \\ 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & -2 \\ 6 & 3 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 5 & -2 & 0 \\ 6 & 3 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 9 & 0 & 3 & 0 \end{bmatrix}$$

are 1, 2, 2, and 2, respectively. In **D**, for example, every 3 × 3 submatrix contains a column of zeros and hence has a vanishing determinant, but the 2 × 2 submatrix $\begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}$ has a nonvanishing determinant, so $r(\mathbf{D}) = 2$. ▮

We may regard the rows of an $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ as n-dimensional vectors, which we call the **row vectors** of $\mathbf{A}$ and which we denote as $\mathbf{r}_1, \dots, \mathbf{r}_m$. Similarly, the columns are m-dimensional vectors, which we call the **column vectors** of $\mathbf{A}$ and which we denote as $\mathbf{c}_1, \dots, \mathbf{c}_n$. Further, we define the vector spaces span $\{\mathbf{r}_1, \dots, \mathbf{r}_m\}$ and span $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ as the **row** and **column spaces** of $\mathbf{A}$, respectively. From the definition of dimension, the dimensions of the row and column spaces are equal to the number of LI row vectors and the number of LI column vectors, respectively.

It will be important to be able to calculate the rank of a given matrix efficiently. With that purpose in mind, we recall the **elementary row operations** defined in Section 8.3:

1. Addition of a multiple of one row to another
   Symbolically: $\mathbf{r}_j \rightarrow \mathbf{r}_j + \alpha \mathbf{r}_k$
2. Multiplication of a row by a nonzero constant
   Symbolically: $\mathbf{r}_j \rightarrow \alpha \mathbf{r}_j$
3. Interchange of two rows
   Symbolically: $\mathbf{r}_j \leftrightarrow \mathbf{r}_k$

Furthermore, we defined (in Section 8.3) two matrices to be **row equivalent** if one can be obtained from the other by finitely many elementary row operations. The following theorem provides an efficient means of calculating the rank of a matrix.

THEOREM 10.5.1 *Elementary Row Operations and Rank*
Row equivalent matrices have the same rank. That is, elementary row operations do not alter the rank of a matrix.

Proof: If matrices $\mathbf{A}$ and $\mathbf{B}$ are row equivalent, then $\mathbf{B}$ can be obtained from $\mathbf{A}$ by a finite number of elementary row operations. It follows that each row vector of $\mathbf{B}$ must be a linear combination of the row vectors of $\mathbf{A}$ so the row space of $\mathbf{B}$ must be a subspace of the row space of $\mathbf{A}$. Similarly, the row space of $\mathbf{A}$ must be a subspace of the row space of $\mathbf{B}$. Thus, the row space of $\mathbf{A}$ is identical to the row space of $\mathbf{B}$, and hence the dimension of the row space of $\mathbf{A}$ (which is the rank of $\mathbf{B}$) must equal the dimension of the row space of $\mathbf{B}$ (which is the rank of $\mathbf{B}$). ▪

EXAMPLE 3. Let

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -3 & 4 \\ 2 & 4 & -2 & 5 \\ 0 & 3 & 1 & 3 \\ 2 & 1 & -3 & -2 \end{bmatrix}. \tag{2}$$

Using elementary row operations,

$$\mathbf{A} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 3 & 1 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & -6 \end{bmatrix} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3}$$

where the operations were as follows: in the first step $\mathbf{r}_2 \to \mathbf{r}_2 + (-1)\mathbf{r}_1$ and $\mathbf{r}_4 \to \mathbf{r}_4 + (-1)\mathbf{r}_1$; in the second step $\mathbf{r}_3 \to \mathbf{r}_3 + (-1)\mathbf{r}_2$; and in the final step $\mathbf{r}_4 \to \mathbf{r}_4 + 3\mathbf{r}_3$. Clearly, the rank of the final matrix is 3 because (deleting the fourth row and third column)

$$\begin{vmatrix} 2 & 1 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{vmatrix} = 12 \neq 0.$$

Thus, by Theorem 10.5.1, $r(\mathbf{A}) = 3$. ▮

The idea, then, is to reduce a given matrix $\mathbf{A}$ to row echelon form by means of elementary row operations.* It can be seen that, in that form, the nonzero rows are LI. In Example 3, for instance, several conclusions follow from (3): $r(\mathbf{A}) = 3$, the number of LI vectors among the rows of $\mathbf{A}$ is 3; the dimension of the row space of $\mathbf{A}$ is 3 and a basis for that row space is given by the vectors $[2, 1, -3, 4]$, $[0, 3, 1, 1]$, and $[0, 0, 0, 2]$.

There is a connection between the rank of a matrix and the linear dependence of a set of vectors, studied in Chapter 9:

THEOREM 10.5.2 *Rank and Linear Dependence*
For any matrix $\mathbf{A}$, the number of LI row vectors is equal to the number of LI column vectors and these, in turn, equal the rank of $\mathbf{A}$.†

Thus, if we wish to determine how many vectors in a given vector set $\{\mathbf{u}_1, \dots, \mathbf{u}_k\}$ are LI we can form a matrix $\mathbf{A}$ with $\mathbf{u}_1, \dots, \mathbf{u}_k$ as the rows (or columns) and then use elementary row operations to determine the rank of $\mathbf{A}$.

EXAMPLE 4. How many LI vectors are contained in $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$, where

$$\mathbf{u}_1 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 1 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} -3 \\ -2 \\ 1 \\ -3 \end{bmatrix}, \quad \mathbf{u}_4 = \begin{bmatrix} 4 \\ 5 \\ 3 \\ -2 \end{bmatrix}?$$

*If we scale the first three rows in the final matrix in (3) by $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{2}$, respectively, so as to begin the nonzero rows with leading ones, then we would say that the resulting matrix is in *reduced* row echelon form.

†For proof of the first part of this theorem, see Theorem 3.5.5 in Steven J. Leon's *Linear Algebra*, 3rd ed. (New York: Macmillan, 1990).

If we construct a matrix having these vectors as columns, then we have the $\mathbf{A}$ matrix (2). Using elementary row operations, we saw, in Example 3, that $r(\mathbf{A}) = 3$. Hence, there are three LI vectors in $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ or, put differently, $\dim[\text{span}\,\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}] = 3$.

COMMENT. The ordering of the columns is immaterial. For instance, we could make $\mathbf{u}_1$ the third column, $\mathbf{u}_2$ the first, $\mathbf{u}_3$ the fourth, and $\mathbf{u}_4$ the second because rank depends upon the zeroness or nonzeroness of determinants whereas the interchanging of columns (or rows) merely changes the *sign* of a determinant. ▮

EXAMPLE 5. *Application to Stoichiometry.* To model the combustion of gasoline in an automobile engine, one can begin by writing down a list of well over 100 simultaneous chemical reactions involving the various hydrocarbons, oxygen, nitrogen, and so on. In turn, these reactions can be modeled by ODE's governing the amount of each chemical species as a function of time, and one can solve the resulting system of ODE's by methods such as those discussed in Chapter 6. It is easy to appreciate that solving around 100 coupled ODE's is a difficult undertaking. Thus, it is important to reduce the list of reactions insofar as possible, and we can do this by eliminating ones that are redundant. For instance, if $A + B \rightarrow C$ and $A + C \rightarrow D$, then a third statement, $2A + B + C \rightarrow C + D$, is redundant in that it is implied by the first two.

To illustrate the reduction process, consider the burning of a mixture of CO, H_2, and CH_4 in a furnace, producing CO, CO_2, and H_2O.* Writing all possible reactions that we can think of gives the list

$$CO + \frac{1}{2}O_2 \rightarrow CO_2, \tag{4a}$$

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O, \tag{4b}$$

$$CH_4 + \frac{3}{2}O_2 \rightarrow CO + 2H_2O, \tag{4c}$$

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O, \tag{4d}$$

where (4c) and (4d) represent the partial and complete combustion of CH_4, respectively. How many of these reactions are independent? It is convenient to re-express them symbolically in the equation format

$$\begin{aligned} CO + \tfrac{1}{2}O_2 - CO_2 &= 0, \\ H_2 + \tfrac{1}{2}O_2 - H_2O &= 0, \\ CH_4 + \tfrac{3}{2}O_2 - CO - 2H_2O &= 0, \\ CH_4 + 2O_2 - CO_2 - 2H_2O &= 0, \end{aligned} \tag{5}$$

where the elements of the coefficient matrix

*This example is discussed by Ben Noble in his book *Applications of Undergraduate Mathematics in Engineering* (New York: Macmillan, 1967). In turn, he notes that the problem was contributed by John Mahoney, Department of Chemical Engineering, West Virginia University, Morgantown, WV.

$$\mathbf{A} = \begin{array}{c} \begin{array}{cccccc} \mathrm{CO} & \mathrm{O_2} & \mathrm{CO_2} & \mathrm{H_2} & \mathrm{H_2O} & \mathrm{CH_4} \end{array} \\ \begin{bmatrix} 1 & \frac{1}{2} & -1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 & -1 & 0 \\ -1 & \frac{3}{2} & 0 & 0 & -2 & 1 \\ 0 & 2 & -1 & 0 & -2 & 1 \end{bmatrix} \end{array} \tag{6}$$

are known as stoichiometric coefficients. To determine a minimum set of independent reactions we reduce $\mathbf{A}$ by elementary row operations and obtain

$$\begin{array}{c} \begin{array}{cccccc} \mathrm{CO} & \mathrm{O_2} & \mathrm{CO_2} & \mathrm{H_2} & \mathrm{H_2O} & \mathrm{CH_4} \end{array} \\ \begin{bmatrix} 1 & \frac{1}{2} & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & -2 & 0 \\ 0 & 0 & -1 & -4 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{7}$$

the rank of which is three. Thus, there are three independent reactions such as the list

$$\begin{aligned} \mathrm{CO} + \tfrac{1}{2}\mathrm{O_2} - \mathrm{CO_2} &= 0, \\ \mathrm{O_2} + 2\mathrm{H_2} - 2\mathrm{H_2O} &= 0, \\ -\mathrm{CO_2} - 4\mathrm{H_2} + 2\mathrm{H_2O} + \mathrm{CH_4} &= 0, \end{aligned} \tag{8}$$

implied by (7). That is, $\mathrm{CO} + \frac{1}{2}\mathrm{O_2} \to \mathrm{CO_2}$, and so on. ■

10.5.2. Application of rank to the system $\mathbf{Ax} = \mathbf{c}$. In Chapter 8 we use the method of Gauss elimination both to solve systems of linear algebraic equations and to study the questions of the existence and uniqueness of solutions. Having developed vector and matrix concepts now, we can return to the important problem $\mathbf{Ax} = \mathbf{c}$ and bring these additional concepts to bear. In doing so, it is convenient to have a representative example to refer to.

EXAMPLE 6. Consider the system $\mathbf{Ax} = \mathbf{c}$ given by

$$\begin{aligned} x_1 - x_2 + x_3 + 3x_4 \qquad\quad + 2x_6 &= 4, \\ x_1 \qquad + 3x_3 + 3x_4 - x_5 + 6x_6 &= 3, \\ 2x_1 - x_2 + 2x_3 + x_4 - x_5 + 7x_6 &= 9, \\ x_1 \qquad + 5x_3 + 8x_4 - x_5 + 7x_6 &= \underline{1}. \end{aligned} \tag{9}$$

Carrying out Gauss elimination by applying elementary row operations to the augmented matrix

$$\mathbf{A}|\mathbf{c} = \begin{bmatrix} 1 & -1 & 1 & 3 & 0 & 2 & 4 \\ 1 & 0 & 3 & 3 & -1 & 6 & 3 \\ 2 & -1 & 2 & 1 & -1 & 7 & 9 \\ 1 & 0 & 5 & 8 & -1 & 7 & 1 \end{bmatrix} \tag{10}$$

gives the row echelon result

$$\begin{bmatrix} 1 & -1 & 1 & 3 & 0 & 2 & 4 \\ 0 & 1 & 2 & 0 & -1 & 4 & -1 \\ 0 & 0 & 2 & 5 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{0} \end{bmatrix} \tag{11}$$

and hence the three-parameter family of solutions

$$\begin{gathered} x_6 = \alpha_1, \qquad x_5 = \alpha_2, \qquad x_4 = \alpha_3, \qquad x_3 = -1 - \tfrac{1}{2}\alpha_1 - \tfrac{5}{2}\alpha_3, \\ x_2 = 1 - 3\alpha_1 + \alpha_2 + 5\alpha_3, \qquad x_1 = 6 - \tfrac{9}{2}\alpha_1 + \alpha_2 + \tfrac{9}{2}\alpha_3, \end{gathered} \tag{12}$$

where the parameters $\alpha_1, \alpha_2, \alpha_3$ are arbitrary.

It is illuminating to express (12) in vector form as

$$\begin{aligned} \mathbf{x} = \begin{bmatrix} 6 - \frac{9}{2}\alpha_1 + \alpha_2 + \frac{9}{2}\alpha_3 \\ 1 - 3\alpha_1 + \alpha_2 + 5\alpha_3 \\ -1 - \frac{1}{2}\alpha_1 - \frac{5}{2}\alpha_3 \\ \alpha_3 \\ \alpha_2 \\ \alpha_1 \end{bmatrix} &= \begin{bmatrix} 6 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \alpha_1 \begin{bmatrix} -\frac{9}{2} \\ -3 \\ -\frac{1}{2} \\ 0 \\ 0 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} \frac{9}{2} \\ 5 \\ -\frac{5}{2} \\ 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3. \end{aligned} \tag{13}$$

Observe that $\mathbf{x}_0$ is a *particular solution* of $\mathbf{A}\mathbf{x} = \mathbf{c}$ (i.e., $\mathbf{A}\mathbf{x}_0 = \mathbf{c}$), and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are *homogeneous solutions* (i.e., $\mathbf{A}\mathbf{x}_1 = \mathbf{0}$, $\mathbf{A}\mathbf{x}_2 = \mathbf{0}$, $\mathbf{A}\mathbf{x}_3 = \mathbf{0}$), by the following reasoning. Since (13) is a solution of $\mathbf{A}\mathbf{x} = \mathbf{c}$ for any α_j's, we can set $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and conclude that $\mathbf{A}\mathbf{x}_0 = \mathbf{c}$. Next, put (13) into $\mathbf{A}\mathbf{x} = \mathbf{c}$:

$$\mathbf{A}\left(\mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3\right) = \mathbf{c}, \tag{14}$$

hence, $\mathbf{A}\mathbf{x}_0 + \alpha_1\mathbf{A}\mathbf{x}_1 + \alpha_2\mathbf{A}\mathbf{x}_2 + \alpha_3\mathbf{A}\mathbf{x}_3 = \mathbf{c}$ or, since $\mathbf{A}\mathbf{x}_0 = \mathbf{c}$,

$$\alpha_1\mathbf{A}\mathbf{x}_1 + \alpha_2\mathbf{A}\mathbf{x}_2 + \alpha_3\mathbf{A}\mathbf{x}_3 = \mathbf{0}. \tag{15}$$

The choice $\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$ reveals that $\mathbf{A}\mathbf{x}_1 = \mathbf{0}$; $\alpha_2 = 1, \alpha_1 = \alpha_3 = 0$ reveals that $\mathbf{A}\mathbf{x}_2 = \mathbf{0}$; and $\alpha_3 = 1, \alpha_1 = \alpha_2 = 0$ reveals that $\mathbf{A}\mathbf{x}_3 = \mathbf{0}$, as claimed.

Observe further that $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are LI, for the rank of

$$[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] = \begin{bmatrix} -\frac{9}{2} & 1 & \frac{9}{2} \\ -3 & 1 & 5 \\ -\frac{1}{2} & 0 & -\frac{5}{2} \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{16}$$

is 3, as is readily seen from the bottom three rows. ▮

Generalizing the results of Example 6, suppose that a system

$$\mathbf{Ax} = \mathbf{c}, \tag{17}$$

where $\mathbf{A}$ is $m \times n$, has a p-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \cdots + \alpha_p \mathbf{x}_p. \tag{18}$$

Then $\mathbf{x}_0$ is necessarily a particular solution, and $\mathbf{x}_1, \dots, \mathbf{x}_p$ are necessarily LI homogeneous solutions. We call span $\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ the **solution space** of the homogeneous equation $\mathbf{Ax} = \mathbf{0}$, or the **null space** of $\mathbf{A}$. The dimension of that null space is called the **nullity** of $\mathbf{A}$.

It is helpful to see that if $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$ denote the columns of $\mathbf{A}$, then $\mathbf{Ax} = \mathbf{c}$ can be expressed as

$$x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_n\mathbf{c}_n = \mathbf{c}, \tag{19}$$

from which we can see that (17) is consistent if and only if $\mathbf{c}$ happens to be in the column space of $\mathbf{A}$ [namely, span $\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$]. Or, in terms of rank, (17) is consistent if and only if the rank of the augmented matrix $\mathbf{A}|\mathbf{c}$ equals the rank of the coefficient matrix $\mathbf{A}$: $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$.

In Example 6, for instance, we see from (11) that $r(\mathbf{A}|\mathbf{c}) = 3$ and (by covering up the last column, which is $\mathbf{c}$) that $r(\mathbf{A}) = 3$ as well. Thus, $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$ and, sure enough, (9) is consistent, solutions being given by (12). However, if we modify (9) by changing the underlined 1 to a 2, say, then the underlined 0 in (11) becomes a 1. In that case $r(\mathbf{A}|\mathbf{c}) = 4$ and $r(\mathbf{A}) = 3$ are unequal and there is no solution because the bottom row of (11) would then be equivalent to $0x_1 + \cdots + 0x_6 = 1$, which cannot be satisfied by any combination of x_j's.

Finally, what can be said about p in (18)? In Example 6, we see from (11) that $p = 3$ for there are three arbitrary x_j values (as seen from the third row), and that that value arises as the difference between the number of unknowns $n = 6$ and the rank $r = 3$.

Let us summarize, for any system $\mathbf{Ax} = \mathbf{c}$.

THEOREM 10.5.3 *Existence and Uniqueness for* $\mathbf{Ax} = \mathbf{c}$
Consider the linear system

$$\mathbf{Ax} = \mathbf{c}, \tag{20}$$

where $\mathbf{A}$ is $m \times n$. There is

1. no solution if and only if $r(\mathbf{A}|\mathbf{c}) \neq r(\mathbf{A})$,
2. a unique solution if and only if $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n$,
3. an $(n - r)$-parameter family of solutions if and only if $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) \equiv r$ is less than n.

The essential ideas required to prove these three results were developed above, so the proofs are left for the exercises.

Naturally, the **homogeneous** system

$$\mathbf{Ax} = \mathbf{0} \tag{21}$$

is but a special case of (20), hence, it is already covered by Theorem 10.5.3. However, it is such an important case that it deserves special attention. In (21), the augmented matrix $r(\mathbf{A}|\mathbf{c})$ is the $\mathbf{A}$ matrix augmented by a column of zeros so it is surely true that $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$ and, according to Theorem 10.5.3, it must be true that (21) is consistent. That result is no great surprise since (21) always admits the trivial solution $\mathbf{x} = \mathbf{0}$. Hence, the significant question about (21) is not whether or not it is consistent, but whether $\mathbf{x} = \mathbf{0}$ is the *only* solution. That is, does (21) admit *non*trivial solutions as well? That question is answered by parts 1 and 2 of Theorem 10.5.3 so we can state the following more specialized results.

THEOREM 10.5.4 *Homogeneous Case Where* $\mathbf{A}$ *is* $m \times n$
If $\mathbf{A}$ is $m \times n$, then

$$\mathbf{Ax} = \mathbf{0} \tag{22}$$

1. is consistent,
2. admits the trivial solution $\mathbf{x} = \mathbf{0}$,
3. admits the unique solution $\mathbf{x} = \mathbf{0}$ if and only if, $r(\mathbf{A}) = n$,
4. admits an $(n - r)$-parameter family of nontrivial solutions, in addition to the trivial solution, if and only if $r(\mathbf{A}) \equiv r < n$.

THEOREM 10.5.5 *Homogeneous Case Where* $\mathbf{A}$ *is* $n \times n$
If $\mathbf{A}$ is $n \times n$, then

$$\mathbf{Ax} = \mathbf{0} \tag{23}$$

admits nontrivial solutions, besides the trivial solution $\mathbf{x} = \mathbf{0}$, if and only if det$\mathbf{A} = 0$.

As a final example, consider an interesting application of these concepts to "dimensional analysis."

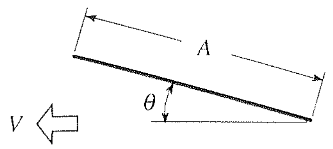


Figure 1. Flat plate airfoil.

EXAMPLE 7. *Dimensional Analysis.* Consider a rectangular flat plate (i.e., a flat rectangular wing, or "airfoil") in steady motion through otherwise-undisturbed air as shown in Fig. 1; V is the flight speed, θ is the incidence or angle of attack of the airfoil, A is the chord length, and B is the span (the dimension normal to the paper). Equivalently, it is experimentally more convenient to keep the airfoil fixed (in a wind tunnel) and to blow air past it, at a speed V. Imagine that our object is to conduct an experimental determination of the lift force ℓ generated on the airfoil, that is, to experimentally determine the functional dependence of ℓ on the various relevant quantities. What quantities *are* relevant? Surely A, B, θ, V are important, as well as the air density ρ (for instance we expect a much greater lift in water than in air, and no lift at all in a vacuum). A reasonable list of the relevant variables is given in Table 1. Other variables come to mind, such as the ambient temperature, but if we expect ℓ to be only weakly dependent on them we can leave them out.

Table 1. Relevant variables.

Variable	Symbol	Fundamental Units
Chord	A	L
Span	B	L
Incidence	θ	$M^0L^0T^0$
Flight velocity	V	LT^{-1}
Velocity of sound in air	V_0	LT^{-1}
Air density	ρ	ML^{-3}
Absolute viscosity	μ	$ML^{-1}T^{-1}$
Lift	ℓ	MLT^{-2}

Major difficulties are now apparent. If ℓ depends upon the seven variables listed in Table 1, and we measure ℓ for five different values of each variable, then we need to conduct $5^7 = 78{,}125$ experimental runs, and then present the results (using graphs, tables, or whatever) in a user-friendly way. Furthermore, whereas some variables are easily varied (such as A, B, θ, V) others are not (such as ρ, μ). The principal object of the following "dimensional analysis" is to reduce the number of variables as much as possible.

To begin, we express each variable in terms of the fundamental units M (mass), L (length), and T (time), in the right-hand column. We do not need to include F (force) as a fundamental unit because, according to Newton's second law, $F = MLT^{-2}$, dimensionally speaking. Also, observe that θ is dimensionless.*

Next, we seek all possible dimensionless products of the form

$$A^aB^b\theta^cV^dV_0^e\rho^f\mu^g\ell^h. \tag{24}$$

That is, we seek the exponents $a, \ldots, h$ such that

$$\begin{aligned}&(L)^a\,(L)^b\,(M^0L^0T^0)^c\,(LT^{-1})^d\,(LT^{-1})^e\,(ML^{-3})^f\,(ML^{-1}T^{-1})^g\,(MLT^{-2})^h\\&= M^0L^0T^0.\end{aligned} \tag{25}$$

Equating exponents of L, T, M on both sides, we see that $a, \ldots, h$ must satisfy the homo-

*Recall that angle is defined by the formula $s = r\theta$, where s is the arc length of a circular arc of radius r, subtended by an angle θ, measured in radians. Thus, $\theta = s/r =$ length / length $=$ dimensionless.

geneous linear system

$$\begin{aligned} a+b+\quad d+e-3f-g+\ h&=0,\\ -d-e\qquad\quad -g-2h&=0,\\ f+g+\ h&=0. \end{aligned} \tag{26}$$

Solving (26) by Gauss elimination gives the five-parameter family of solutions

$$\begin{bmatrix} a\\ b\\ c\\ d\\ e\\ f\\ g\\ h \end{bmatrix} = \alpha_1 \begin{bmatrix} -2\\ 0\\ 0\\ -2\\ 0\\ -1\\ 0\\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} -1\\ 0\\ 0\\ -1\\ 0\\ -1\\ 1\\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} 0\\ 0\\ 0\\ -1\\ 1\\ 0\\ 0\\ 0 \end{bmatrix} + \alpha_4 \begin{bmatrix} 0\\ 0\\ 1\\ 0\\ 0\\ 0\\ 0\\ 0 \end{bmatrix} + \alpha_5 \begin{bmatrix} -1\\ 1\\ 0\\ 0\\ 0\\ 0\\ 0\\ 0 \end{bmatrix}, \tag{27}$$

where $\alpha_1, \dots, \alpha_5$ are arbitrary constants. With $\alpha_1 = 1$ and $\alpha_2 = \cdots = \alpha_5 = 0$, say, (27) gives $a = -2, b = c = 0$, etc., and hence the nondimensional parameter $A^{-2}B^0\theta^0V^{-2}V_0^0\rho^{-1}\mu^0\ell^1$, namely, the nondimensional lift $\ell/(\rho V^2A^2)$.* Similarly, setting $\alpha_2 = 1$ and the other α_j's $= 0$ gives the nondimensional parameter $\rho AV/\mu$, well known in fluid mechanics as the *Reynolds number* and denoted as Re; setting $\alpha_3 = -1$ and the other α_j's $= 0$ gives the *Mach number* V/V_0, denoted as $\mathcal{M}$; setting $\alpha_4 = 1$ and the other α_j's $= 0$ gives the *incidence* θ (which was nondimensional to begin with); and setting $\alpha_5 = 1$ and the other α_j's $= 0$ gives B/A, known as the *aspect ratio*, denoted typically as $\mathcal{AR}$.

The upshot is that rather than seek a functional relationship on the eight variables listed in the table, we can seek a relationship on the five nondimensional variables $[\ell/(\rho V^2A^2), \text{Re}, \mathcal{M}, \theta, \mathcal{AR}]$. Or, singling out the nondimensional lift, we can express

$$\frac{\ell}{\rho V^2A^2} = f\,(\text{Re}, \mathcal{M}, \theta, \mathcal{AR}) \tag{28}$$

and determine f experimentally by measuring $\ell/(\rho V^2A^2)$ for various combinations of Re, $\mathcal{M}$, θ, and $\mathcal{AR}$ values.

COMMENT 1. A trained fluid mechanicist could probably simplify the problem even further. For example, it is known (from the governing equations of fluid mechanics) that the effect of the Mach number $\mathcal{M}$ will be negligible if $\mathcal{M}^2 \ll 1$. Thus, if we have flight speeds V less than 300 miles per hour in mind, then $\mathcal{M}$ can, to a good approximation, be dropped from (28).†

COMMENT 2. We mentioned, above, the practical difficulty in carrying out the experiment for a range of values of the fluid density. For instance, we could use air and water, but

*A fluid mechanicist would probably change this to $\ell/(\rho V^2AB)$ (corresponding to $\alpha_1 = 1, \alpha_2 = \alpha_3 = \alpha_4 = 0, \alpha_5 = -1$), or to $\ell/(\frac{1}{2}\rho V^2AB)$ because $\frac{1}{2}\rho V^2$ has physical significance as the stagnation pressure, and AB is the area of the airfoil.

†At ground level, the speed of sound is 762 mph, so if $V = 300$ mph then $\mathcal{M} = (300/762)^2 = 0.155$ is indeed small compared to 1.

their densities are widely different and we would need both wind tunnel and water tunnel facilities. In the right-hand side of (28), however, ρ shows up only within the Reynolds number $\text{Re} = \rho AV/\mu$, which can be varied readily by varying the wind speed V.

COMMENT 3. Of course, (27) gives an *infinite* number of nondimensional parameters. However, there are only five independent ones, such as the ones named above. For instance, we could choose $\alpha_3 = \alpha_5 = 1$ and $\alpha_1 = \alpha_2 = \alpha_4 = 0$, but the resulting nondimensional parameter, $V_0B/(VA)$, is merely the aspect ratio divided by the Mach number. ∎

Closure. The rank $r(\mathbf{A})$ is defined as the size of the largest nonvanishing determinant within $\mathbf{A}$. Because the rank of a matrix is unaffected by elementary row operations, we can determine the rank of a given matrix efficiently by reducing it to row echelon form, in which form the rank can be seen by inspection. Principal applications of the concept of rank include the calculation of the number of LI vectors (n-tuple vectors, that is) within a given set, and the theory of the existence and uniqueness of solutions of systems of linear algebraic equations.

Computer software. Using *Maple*, we can evaluate the rank of a given matrix using the **rank(A)** command. For instance, to evaluate $r(\mathbf{A})$ where the rows of $\mathbf{A}$ are $[1, 2, 3, 4]$, $[2, 4, 6, 8]$, and $[1, 1, 1, 1]$, respectively, enter

```
with(linalg):
```

to access the rank(A) command. Then enter

```
rank(array([[1, 2, 3, 4, ], [2, 4, 6, 8], [1, 1, 1, 1]]));
```

and return. The output is 2. Alternatively, the sequence

```
A := array([[1, 2, 3, 4, ], [2, 4, 6, 8], [1, 1, 1, 1]]):
rank(A);
```

gives the same result.

EXERCISES 10.5

1. Determine the rank, nullity, number of LI rows, and number of LI columns for the given matrix.

(a) $[0, 0, 2, 0]$

(b) $[1, 2, 3]$

(c) $\begin{bmatrix} 5 & 7 \\ 4 & 9 \end{bmatrix}$

(d) $\begin{bmatrix} 4 & 8 & 0 \\ 3 & 6 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$

(f) $\begin{bmatrix} 3 & 2 & 1 \\ -1 & 1 & 4 \\ 1 & 4 & 9 \end{bmatrix}$

(g) $\begin{bmatrix} 5 & 0 & 0 \\ 3 & 0 & 4 \\ 2 & 0 & 0 \end{bmatrix}$

(h) $\begin{bmatrix} 1 & 3 & 2 \\ 2 & 6 & 4 \\ 3 & 9 & 6 \end{bmatrix}$

(i) $\begin{bmatrix} 6 & 5 & 2 & 0 \\ 0 & 2 & 3 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}$

(j) $\begin{bmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \\ 1 & 0 & 3 & 1 \end{bmatrix}$

(k) $\begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & -1 \\ 0 & 1 & 3 \\ 1 & 2 & 4 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 4 & -1 & 1 \\ 1 & 1 & 5 & -1 \\ 1 & 5 & 4 & 0 \\ 2 & 6 & 9 & -1 \end{bmatrix}$

(m) $\begin{bmatrix} 3 & 1 & -4 & 4 \\ 0 & 2 & 2 & 9 \\ 2 & 2 & 1 & 0 \\ -1 & 3 & 7 & 5 \end{bmatrix}$ (n) $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & -1 & 0 & 5 \\ 6 & -1 & 2 & 8 \\ 7 & -1 & 0 & 7 \\ 8 & -1 & -3 & 5 \end{bmatrix}$

2. (a)–(n) Use computer software to determine the rank of the matrix given in the corresponding part of Exercise 1.

3. (a)–(n) Consider the problem $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is the $m \times n$ matrix given in the corresponding part of Exercise 1. In each case let $\mathbf{c}$ be the m-dimensional vector $[1, 1, \dots, 1]^T$. Use Theorem 10.5.3, and suitable rank calculations, to determine whether or not the system is consistent. If consistent, determine whether it admits a unique solution or a p-parameter family of solutions. If the latter, determine p. Do not solve the system; merely use the concept of rank and Theorem 10.5.3.

4. If two matrices of the same form have the same rank, need they be row equivalent? Prove or disprove.

5. Show, by carrying out suitable row operations, that the following pairs of matrices are row equivalent.

(a) $\begin{bmatrix} 1 & 3 & -3 & 0 \\ 2 & 1 & 0 & 4 \end{bmatrix}$ and $\begin{bmatrix} 4 & 2 & 0 & 8 \\ 3 & 4 & -3 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & 1 & 0 \\ 2 & 2 & 3 \\ 4 & 4 & 9 \\ 4 & 4 & 6 \end{bmatrix}$ and $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 5 & 2 \\ 0 & -3 \\ 1 & 4 \end{bmatrix}$ and $\begin{bmatrix} 9 & 2 \\ -6 & -3 \\ 9 & 4 \end{bmatrix}$

6. Show that the following pairs of matrices are *not* row equivalent.

(a) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 4 \\ 6 & 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 3 & -1 & 5 \\ 1 & 5 & -5 \\ 2 & 2 & 0 \end{bmatrix}$

7. Exercises 5 and 6 are simple enough to be solvable by inspection. More generally, inspection may not suffice. Put forward a systematic procedure for determining whether or not two given matrices are row equivalent, and apply that procedure to the given matrices.

(a) $\begin{bmatrix} 1 & 2 & 3 & -1 \\ 2 & 4 & -1 & 1 \\ 0 & 5 & 6 & 3 \\ 4 & -2 & 0 & 6 \end{bmatrix}$ and $\begin{bmatrix} 1 & -1 & 2 & 1 \\ 3 & 0 & 5 & 1 \\ 2 & 2 & 1 & 3 \\ 3 & 1 & 3 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 4 & -1 & 2 & 3 \\ 0 & 1 & 1 & 2 \\ -1 & 0 & 2 & 0 \end{bmatrix}$ and $\begin{bmatrix} 3 & 0 & 5 & 5 \\ 6 & 0 & -1 & 5 \\ 2 & 0 & 7 & 5 \end{bmatrix}$

8. Determine whether the following set of vectors is LI or LD by computing the rank of a suitable matrix and invoking the relevant theorem.

(a) $[2, 0, 1, -1]$, $[0, 3, 0, 3]$, $[4, 3, 2, 1]$
(b) $[4, 1, 2]$, $[2, 2, 1]$, $[2, -1, 1]$, $[4, 7, 2]$, $[0, 1, 0]$
(c) $[1, 3, 2, 4, 5]$, $[2, 3, 1, 5, 4]$, $[4, 5, 3, 1, 2]$
(d) $[2, 1, 1]$, $[4, 2, 2]$, $[0, 1, 2]$, $[1, 0, 0]$
(e) $[1, -2, 0, 1]$, $[0, 1, 1, 2]$, $[1, 0, 2, 5]$, $[2, -7, -3, -4]$

9. Prove that $r(\mathbf{A}^{\mathrm{T}}) = r(\mathbf{A})$ for every $m \times n$ matrix using any results given in this section.

10. The property $\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B})$, of determinants, where $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$, might seem to imply that $r(\mathbf{AB}) = r(\mathbf{A})r(\mathbf{A})$. Is the latter true? Prove or disprove.

11. Is $r(\mathbf{A} + \mathbf{B}) = r(\mathbf{A}) + r(\mathbf{B})$ true? Prove or disprove.

12. Prove that if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then $r(\mathbf{AB}) \leq n$. HINT: Partition $\mathbf{B}$ into rows, and write

$$\mathbf{AB} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_n \end{bmatrix} = \begin{bmatrix} a_{11}\mathbf{r}_1 + \cdots + a_{1n}\mathbf{r}_n \\ \vdots \\ a_{m1}\mathbf{r}_1 + \cdots + a_{mn}\mathbf{r}_n \end{bmatrix}.$$

13. (a) Below (18), we stated that $\mathbf{x}_1, \dots, \mathbf{x}_p$ are necessarily LI. Prove that claim. HINT: Pattern your proof after the discussion in Example 6.
(b) Show that $\mathbf{x}_0$ cannot be in the span of $\mathbf{x}_1, \dots, \mathbf{x}_p$.

14. Although we made a case for the truth of Theorem 10.5.3, we did not provide a detailed proof.

(a) Prove part 1. (b) Prove part 2. (c) Prove part 3.

15. This exercise refers to Example 6 and the discussion following that example. For definiteness, let $\mathbf{A}$ be 3×3.

(a) Suppose that $\mathbf{Ax} = \mathbf{c}$ admits a one-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1. \tag{15.1}$$

Explain, with the help of a labeled sketch, the geometrical significance of $\mathbf{x}_0, \mathbf{x}_1$ and $\mathbf{x}_0 + \alpha_1 \mathbf{x}_1$.
(b) Suppose that $\mathbf{Ax} = \mathbf{c}$ admits a two-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2. \tag{15.2}$$

Explain, with the help of a labeled sketch, the geometrical significance of $\mathbf{x}_0$ and $\mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2$.

16. (*Stoichiometry*) Determine how many of the following reactions are independent, and give such a set of independent reactions.

(a) $H_2 + O_2 \rightleftharpoons 2OH$,
$H_2 + \frac{1}{2}O_2 \rightleftharpoons H_2O$,
$H + OH \rightleftharpoons H_2 + O$,
$H_2 \rightleftharpoons 2H$,
$O_2 \rightleftharpoons 2O$

(b) $H_2 + Cl_2 \rightleftharpoons 2HCl$,
$Cl + H_2 \rightleftharpoons HCl + H$,
$H_2 \rightleftharpoons 2H$,
$Cl_2 \rightleftharpoons 2Cl$

(c) $O_2 + CH_3 \rightarrow CH_3OO$,
$CH_4 + CH_3OO \rightarrow CH_3 + CH_3OOH$,
$CH_3OOH \rightarrow CO + 2H_2 + O$,
$CH_4 + O \rightarrow CH_3 + OH$,
$CH_4 + O_2 \rightarrow CH_3OOH$

17. (*Dimensional analysis*) In studying the drag force on a sphere moving beneath a water surface, the tabulated variables are deemed relevant. Proceeding along the same lines as in Example 7, obtain the following relevant dimensionless parameters: the dimensionless drag force $D/(\rho V^2 R^2)$, the *Reynolds number* Re $= \rho RV/\mu$, the *Froude number* $V^2/(Rg)$, and the two length ratios λ/R and d/R.

Variable	Symbol	Fundamental Units
Radius of sphere	R	L
Depth below water surface	d	L
Velocity of sphere	V	LT^{-1}
Water density	ρ	ML^{-3}
Absolute viscosity	μ	$ML^{-1}T^{-1}$
Gravity	g	LT^{-2}
Wavelength of free surface waves	λ	L
Drag force	D	MLT^{-2}

10.6 Inverse Matrix, Cramer's Rule, Factorization

There exist important methods for solving a linear system $\mathbf{Ax} = \mathbf{c}$ besides Gauss elimination. In this section we study three: the inverse matrix method, Cramer's rule, and LU factorization.

10.6.1. Inverse matrix. Having introduced matrix notation so that a system of linear algebraic equations can be expressed compactly as

$$\mathbf{Ax} = \mathbf{c}, \tag{1}$$

the form of (1) itself suggests other solution strategies. For how would we solve the simple scalar equation $3x = 12$? We could divide both sides by 3 and obtain $x = \frac{12}{3} = 4$. However, if we try to carry that idea over to the matrix case (1), we obtain $\mathbf{x} = \dfrac{\mathbf{c}}{\mathbf{A}}$, and need to know how to divide one matrix into another. However, matrix division has not been (and will not be) defined. Alternatively, we can solve

$3x = 12$ by multiplying both sides by $\frac{1}{3}$, for that step gives $\frac{1}{3}3x = \frac{1}{3}12$, or $1x = 4$, and hence $x = 4$. That idea does carry over to (1) because matrix *multiplication* is defined.

The idea, then, is to seek a matrix "$\mathbf{A}^{-1}$" having the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for then

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{2}$$

becomes

$$\mathbf{I}\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}, \tag{3}$$

and since $\mathbf{I}\mathbf{x} = \mathbf{x}$, we have the solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{4}$$

of (1). Note that $\mathbf{A}^{-1}$ does not mean $1/\mathbf{A}$ or $\mathbf{I}/\mathbf{A}$; it is simply the name of the matrix having the property

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \tag{5}$$

if one exists. We call it the **inverse** of $\mathbf{A}$, or "$\mathbf{A}$-inverse" for brevity.

Consider an exploratory example.

EXAMPLE 1. Let $\mathbf{A}\mathbf{x} = \mathbf{c}$ be the system

$$\begin{aligned} -x_1 - 2x_2 &= 1 \\ x_1 + x_2 &= 1 \\ x_2 &= 1 \end{aligned} \quad \text{or} \quad \begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{6}$$

We find, for example by trial and error, that

$$\begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}, \tag{7}$$

so that the pre-multiplication of (6) by the matrix $\begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$ yields

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \tag{8}$$

and hence the (tentative) solution $x_1 = 3, x_2 = 4$. Yet the latter does not satisfy (6), so the method is in some way incorrect. [In fact, Gauss elimination reveals that (6) is inconsistent, has no solution.] ▮

In other words, we must proceed with caution. The idea is that the pre-multiplication of $\mathbf{A}\mathbf{x} = \mathbf{c}$ by a matrix $\mathbf{A}^{-1}$ having the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ *does not necessarily lead to an equivalent system.*

However, we now show that if $\mathbf{A}$ is square, say $n \times n$, and a matrix $\mathbf{A}^{-1}$ can be found such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, then the pre-multiplication of $\mathbf{Ax} = \mathbf{c}$ by $\mathbf{A}^{-1}$ *does* lead to an equivalent system, namely, the unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.

First, assuming that $\mathbf{A}$ is $n \times n$, observe from

$$\underbrace{\mathbf{A}^{-1}}_{m\times p}\;\underbrace{\mathbf{A}}_{n\times n} = \mathbf{I} \tag{9}$$

that p (the number of columns in $\mathbf{A}^{-1}$) must equal n for $\mathbf{A}^{-1}$ to be conformable for multiplication with $\mathbf{A}$, and that m (the number of rows in $\mathbf{A}^{-1}$) must equal n for the product $\mathbf{A}^{-1}\mathbf{A}$ to be square. Thus, $\mathbf{A}^{-1}$ is necessarily square too, of order n.

Second, it follows from (9) that

$$\det(\mathbf{A}^{-1}\mathbf{A}) = \det\mathbf{I} = 1$$

or, since the determinant of a product equals the product of the determinants (property D12 in Section 10.4),

$$(\det\mathbf{A}^{-1})(\det\mathbf{A}) = 1. \tag{10}$$

Equation (10) cannot possibly be satisfied if $\det\mathbf{A} = 0$. Hence, if a matrix $\mathbf{A}^{-1}$ satisfying (9) is to exist, it is *necessary* that $\det\mathbf{A} \neq 0$.

Assuming that that is the case, that $\det\mathbf{A} \neq 0$, let us seek to determine $\mathbf{A}^{-1}$. Our starting point is the cofactor expansion from Section 10.4,

$$\det\mathbf{A} = \sum a_{jk}A_{jk}, \tag{11}$$

where A_{jk} is the cofactor of the a_{jk} element, and the sum is either on j (for any fixed value of k) or on k (for any fixed value of j). Let us take the sum to be on j. Observe that

$$\sum_j a_{jk}A_{ji} = \begin{cases} \det\mathbf{A} & \text{if } i = k, \\ 0 & \text{if } i \neq k \end{cases} \tag{12}$$

since if $i = k$, then (11) applies, and if $i \neq k$, then the left-hand side of (12) is again a cofactor expansion, this time an expansion about the ith column – but with the ith column replaced by the kth column; thus, it is a determinant containing two identical columns and according to property D5 in Section 10.4, it must therefore be zero. Rearranging (12) by dividing through by $\det\mathbf{A}$ (which is permissible since we have assumed that $\det\mathbf{A} \neq 0$) and using the Kronecker delta notation,*

$$\sum_j \left(\frac{A_{ji}}{\det\mathbf{A}}\right) a_{jk} = \delta_{ik}. \tag{13}$$

*Defined in Section 9.10.2, δ_{ik} is simply 1 if $i = k$ and 0 if $i \neq k$. Thus, a matrix $\{\delta_{ik}\}$ is an identity matrix $\mathbf{I}$.

This scalar statement (which holds for $1 \leq i \leq n$ and $1 \leq j \leq n$) is equivalent, according to the definition of matrix multiplication,[†] to the matrix equation

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \tag{14}$$

where the desired **inverse matrix** $\mathbf{A}^{-1}$ is[‡]

$$\mathbf{A}^{-1} = \{\alpha_{ij}\} = \left\{ \frac{A_{ji}}{\det\mathbf{A}} \right\}. \tag{15}$$

Or, written out,

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}} \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}. \tag{16}$$

CAUTION: The A_{ji} in (15) is not a misprint; the j, i indices simply turn out to be in the reverse order of those in α_{ij}. For instance, the $1, 2$ element of $\mathbf{A}^{-1}$ is $A_{21}/\det\mathbf{A}$, not $A_{12}/\det\mathbf{A}$, where A_{21} is the *cofactor* of the $2, 1$ element in $\mathbf{A}$, not the $2, 1$ *element* of $\mathbf{A}$ (which is a_{21}).

The matrix in (16) is called the **adjoint of A** and is denoted as adj**A** so

$$\boxed{\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}} \operatorname{adj}\mathbf{A}.} \tag{17}$$

To form the adjoint of a given square matrix **A** we replace each element by its cofactor, and take the transpose of the resulting matrix.

The upshot, then, is that if $\det\mathbf{A} \neq 0$, then $\mathbf{A}^{-1}$ exists and is given by (17). In that case we say that **A** is **invertible**. If $\det\mathbf{A} = 0$, then $\mathbf{A}^{-1}$ does not exist, and we say that **A** is **singular**.

EXAMPLE 2. Determine the inverse of

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ 0 & 1 & 4 \\ 1 & 5 & -2 \end{bmatrix}, \tag{18}$$

if it exists. It does exist because $\det\mathbf{A} = -57 \neq 0$ and is given by (17). Since adj**A** is the transpose of the cofactor matrix, we have

$$\operatorname{adj}\mathbf{A} = \begin{bmatrix} \begin{vmatrix} 1 & 4 \\ 5 & -2 \end{vmatrix} & -\begin{vmatrix} 0 & 4 \\ 1 & -2 \end{vmatrix} & \begin{vmatrix} 0 & 1 \\ 1 & 5 \end{vmatrix} \\ -\begin{vmatrix} 2 & -1 \\ 5 & -2 \end{vmatrix} & \begin{vmatrix} 3 & -1 \\ 1 & -2 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ 1 & 5 \end{vmatrix} \\ \begin{vmatrix} 2 & -1 \\ 1 & 4 \end{vmatrix} & -\begin{vmatrix} 3 & -1 \\ 0 & 4 \end{vmatrix} & \begin{vmatrix} 3 & 2 \\ 0 & 1 \end{vmatrix} \end{bmatrix}^{\mathrm{T}}$$

[†] Recall that if $\mathbf{BC} = \mathbf{D}$, then $d_{ij} = \sum_k b_{ik}c_{kj}$ or, what is equivalent, $d_{ik} = \sum_j b_{ij}c_{jk}$.

[‡] It is tempting to let the i, j element of $\mathbf{A}^{-1}$ be denoted as a_{ij}^{-1}, but this quantity could be misunderstood to be $1/a_{ij}$. Thus, let us use α_{ij}.

$$= \begin{bmatrix} -22 & 4 & -1 \\ -1 & -5 & -13 \\ 9 & -12 & 3 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} -22 & -1 & 9 \\ 4 & -5 & -12 \\ -1 & -13 & 3 \end{bmatrix} \tag{19}$$

so

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\,\mathrm{adj}\mathbf{A} = -\frac{1}{57}\begin{bmatrix} -22 & -1 & 9 \\ 4 & -5 & -12 \\ -1 & -13 & 3 \end{bmatrix}. \tag{20}$$

It is readily verified from (20) and (18) that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ▮

Besides having $\mathbf{A}^{-1}$ satisfy $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, as stated in (14), it is crucial, as we shall see, to have $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ as well. To show that $\mathbf{A}\mathbf{A}^{-1}$ does equal $\mathbf{I}$, write

$$\mathbf{A}\mathbf{A}^{-1} = \left\{\sum_k a_{ik}\alpha_{kj}\right\} = \left\{\sum_k a_{ik}\frac{A_{jk}}{\det\mathbf{A}}\right\} = \frac{1}{\det\mathbf{A}}\left\{\sum_k a_{ik}A_{jk}\right\}, \tag{21}$$

where the first equality follows from the definition of matrix multiplication, and the second follows from (15). Now,

$$\sum_k a_{ik}A_{jk} = \begin{cases} \det\mathbf{A}, & i = j \\ 0, & i \neq j \end{cases} \tag{22}$$

because if $i = j$ then (11) applies, and if $i \neq j$ then the left-hand side of (22) is again a cofactor expansion, this time an expansion about the jth row – but with the jth row replaced by the ith row. Thus, it is a determinant containing two identical rows and, according to property D5, it must therefore be zero. Hence (21) becomes

$$\mathbf{A}\mathbf{A}^{-1} = \{\delta_{ij}\} = \mathbf{I}, \tag{23}$$

as claimed. Thus,

$$\boxed{\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.} \tag{24}$$

In view of the first equality in (24), we see that $\mathbf{A}^{-1}$ and $\mathbf{A}$ necessarily commute.

To understand the significance of (24), let us review the solution of $\mathbf{A}\mathbf{x} = \mathbf{c}$ by the inverse matrix method. The steps (2) and (3) gave $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$, provided that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. To verify that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ does indeed satisfy $\mathbf{A}\mathbf{x} = \mathbf{c}$, let us put $\mathbf{A}^{-1}\mathbf{c}$ into that equation in place of $\mathbf{x}$:

$$\mathbf{A}\left(\mathbf{A}^{-1}\mathbf{c}\right) = \mathbf{c} \tag{25}$$

or, by the associative property of matrix multiplication,

$$\left(\mathbf{A}\mathbf{A}^{-1}\right)\mathbf{c} = \mathbf{c}, \tag{26}$$

which is indeed true because $\mathbf{AA}^{-1} = \mathbf{I}$.*

Let us pull these results together.

THEOREM 10.6.1 *Inverse Matrix*
Let $\mathbf{A}$ be $n \times n$. If $\det\mathbf{A} \neq 0$, then there exists a unique matrix $\mathbf{A}^{-1}$, also $n \times n$, called the inverse of $\mathbf{A}$, such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}. \tag{27}$$

$\mathbf{A}$ is then said to be invertible, and its inverse is given by (17). If $\det\mathbf{A} = 0$, then a matrix $\mathbf{A}^{-1}$ satisfying (27) does not exist, and $\mathbf{A}$ is said to be singular.*

Proof: In the discussion preceding the theorem we proved all but the uniqueness of $\mathbf{A}^{-1}$. To prove uniqueness, let $\mathbf{B}$ and $\mathbf{C}$ both be inverses of $\mathbf{A}$. Then $\mathbf{BA} = \mathbf{I}$ and $\mathbf{CA} = \mathbf{I}$. Subtracting, $\mathbf{BA} - \mathbf{CA} = \mathbf{0}$ or $(\mathbf{B} - \mathbf{C})\mathbf{A} = \mathbf{0}$. And post-multiplying this last equation by $\mathbf{A}^{-1}$ (which exists by assumption), we have $(\mathbf{B} - \mathbf{C})\mathbf{AA}^{-1} = \mathbf{0A}^{-1}$ or $(\mathbf{B} - \mathbf{C})\mathbf{I} = \mathbf{0}$. Thus, $\mathbf{B} - \mathbf{C} = \mathbf{0}$, and hence $\mathbf{B} = \mathbf{C}$. ■

Finally, we return to the application of $\mathbf{A}^{-1}$ in the solution of $\mathbf{Ax} = \mathbf{c}$.

THEOREM 10.6.2 *Solution of* $\mathbf{Ax} = \mathbf{c}$
If $\mathbf{A}$ is $n \times n$ and $\det\mathbf{A} \neq 0$, then $\mathbf{Ax} = \mathbf{c}$ admits the unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.

Proof: That $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ satisfies $\mathbf{Ax} = \mathbf{c}$ was shown just above Theorem 10.6.1. That the solution is unique follows from Theorem 10.5.3 because $\det\mathbf{A} \neq 0$ implies that $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n$. ■

There are several useful properties of inverse matrices.

*Now we can understand that the failure in Example 1 occurred because

$$\begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} -3 & -4 & -4 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix} \neq \mathbf{I}.$$

*There exist interesting generalizations of the notion of the inverse matrix for matrices that are not strictly invertible (perhaps not even square). Such are the *Moore–Penrose generalized inverse* and the *pseudoinverse*. See, for example, Gilbert Strang, *Linear Algebra and Its Applications* (New York: Academic Press, 1976), Chap. 3.

PROPERTIES OF INVERSES

I1. If $\mathbf{A}$ and $\mathbf{B}$ are of the same order, and invertible, then $\mathbf{AB}$ is too, and

$$\boxed{(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.} \tag{28}$$

I2. If $\mathbf{A}$ is invertible, then

$$\boxed{\left(\mathbf{A}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}} \tag{29}$$

and

$$\det\left(\mathbf{A}^{-1}\right) = \frac{1}{\det\mathbf{A}}. \tag{30}$$

I3. If $\mathbf{A}$ is invertible, then $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$ and $(\mathbf{A}^m)^n = \mathbf{A}^{mn}$ for any (positive, negative, or zero) integers m and n.

I4. If $\mathbf{A}$ is invertible, then $\mathbf{AB} = \mathbf{AC}$ implies that $\mathbf{B} = \mathbf{C}$, $\mathbf{BA} = \mathbf{CA}$ implies that $\mathbf{B} = \mathbf{C}$, $\mathbf{AB} = \mathbf{0}$ implies that $\mathbf{B} = \mathbf{0}$, and $\mathbf{BA} = \mathbf{0}$ implies that $\mathbf{B} = \mathbf{0}$.

Of these, let us prove (28) and (29) and leave the remaining proofs as exercises. First (28): Since $\mathbf{A}$ and $\mathbf{B}$ are invertible, by assumption, $\det\mathbf{A} \neq 0$ and $\det\mathbf{B} \neq 0$. Thus, $\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B}) \neq 0$ so $\mathbf{AB}$ is invertible too. Let us denote $(\mathbf{AB})^{-1}$ as $\mathbf{C}$. Then $\mathbf{ABC} = \mathbf{I}$, $\mathbf{A}^{-1}\mathbf{ABC} = \mathbf{A}^{-1}\mathbf{I}$, $\mathbf{BC} = \mathbf{A}^{-1}$, $\mathbf{B}^{-1}\mathbf{BC} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, hence, $\mathbf{C} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, as claimed. As a mnemonic device, note the resemblence of (28) to the transpose formula $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$.

To prove (29), begin with $\left(\mathbf{AA}^{-1}\right)^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}$. But $\left(\mathbf{AA}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$. Hence, $\left(\mathbf{A}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\right)^{-1}$.

10.6.2. Application to a mass-spring system. To illustrate a number of these ideas with a physical application, consider the arrangement of masses and springs shown in Fig. 1. The three masses are in static equilibrium under the action of prescribed applied forces f_1, f_2, f_3, and the k's denote the stiffnesses of the various springs. For instance, k_{12} denotes the stiffness of the spring connecting mass number 1 and mass number 2. Mass-spring systems are discussed in Section 1.3,

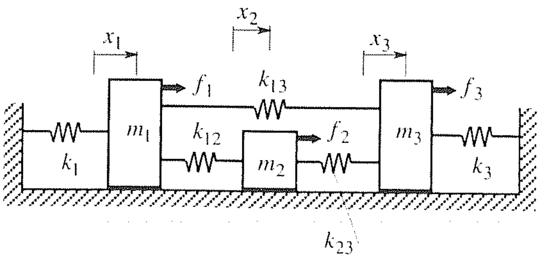


Figure 1. Mass-spring system.

and in Example 3 of Section 3.9.1, which discussions should be reviewed if the following is not clear.

The free-body diagrams (i.e., the force diagrams) of the three masses are shown in Fig. 2, where it has been assumed, simply for definiteness, that $x_1 > x_2 > x_3 > 0$.

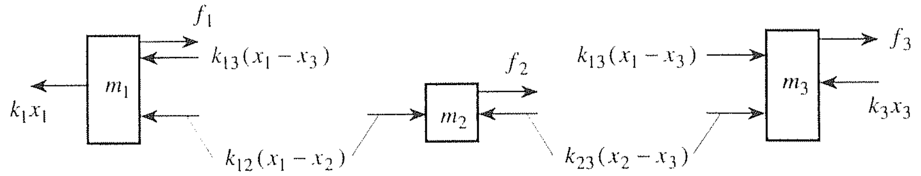


Figure 2. Free-body diagrams.

From that assumption it follows that each spring is in compression except for the left-hand spring of stiffness k_1.

From Fig. 2 and Newton's second law, we obtain the equations of motion

$$\begin{aligned} m_1 x_1'' &= f_1 - k_1 x_1 - k_{12}(x_1 - x_2) - k_{13}(x_1 - x_3), \\ m_2 x_2'' &= f_2 + k_{12}(x_1 - x_2) - k_{23}(x_2 - x_3), \\ m_3 x_3'' &= f_3 + k_{13}(x_1 - x_3) + k_{23}(x_2 - x_3) - k_3 x_3, \end{aligned} \tag{31}$$

where primes denote differentiation with respect to the time t. Since the system is in static equilibrium by assumption, $x_1'' = x_2'' = x_3'' = 0$, and (31) becomes, in matrix form,

$$\begin{bmatrix} (k_1 + k_{12} + k_{13}) & -k_{12} & -k_{13} \\ -k_{12} & (k_{12} + k_{23}) & -k_{23} \\ -k_{13} & -k_{23} & (k_{13} + k_{23} + k_3) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \tag{32}$$

or

$$\mathbf{Kx} = \mathbf{f}, \tag{33}$$

where we will call $\mathbf{K}$ the *stiffness matrix*. We see that (33) is a matrix generalization of the simple Hooke's law $f = kx$ for a single spring.

Is there a unique solution $\mathbf{x} = \mathbf{K}^{-1}\mathbf{f}$? Experience and physical intuition probably tells us there is. Mathematically, everything hinges upon $\det\mathbf{K}$. If $\det\mathbf{K} \neq 0$, there is a unique solution for $\mathbf{x}$, and if $\det\mathbf{K} = 0$, then there is either no solution or an infinity of solutions. With five parameters within $\mathbf{K}$ (namely, $k_1, k_{12}, k_{13}, k_{23}, k_3$) it is hard to imagine that we cannot have $\det\mathbf{K} = 0$ for some choice(s) of those parameters. Let us see. Working out the determinant of $\mathbf{K}$, we find that

$$\begin{aligned} \det\mathbf{K} = k_1 &(k_{12}k_{13} + k_{12}k_{23} + k_{12}k_3 + k_{23}k_{13} + k_{23}k_3) \\ &+ k_3 (k_{12}k_{23} + k_{12}k_{13}) . \end{aligned} \tag{34}$$

Since each sign is positive, and the k's are positive, we see that $\det\mathbf{K} \neq 0$ so there is indeed a unique solution for $\mathbf{x}$, namely, $\mathbf{x} = \mathbf{K}^{-1}\mathbf{f}$.

However, suppose we degrade the system by removing one or more springs. We can see from (34) that even if we set any one k value equal to zero (i.e., remove that spring), $\det\mathbf{K}$ is still positive. If we are willing to remove two springs, then we can obtain $\det\mathbf{K} = 0$ in either of two ways, by setting $k_1 = k_3 = 0$ or by setting $k_{12} = k_{23} = 0$. Let us consider the former, and leave the latter for the exercises.

With $k_1 = k_3 = 0$, $\mathbf{K}$ is singular (noninvertible) and (33) admits either no solution or an infinity of them. Which is it, and how is the result to be understood physically? Setting $k_1 = k_3 = 0$, (32) reduces to

$$\begin{bmatrix} (k_{12}+k_{13}) & -k_{12} & -k_{13} \\ -k_{12} & (k_{12}+k_{23}) & -k_{23} \\ -k_{13} & -k_{23} & (k_{13}+k_{23}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \tag{35}$$

and, in augmented matrix form, Gauss elimination gives

$$\begin{bmatrix} 1 & \dfrac{k_{23}}{k_{13}} & -1-\dfrac{k_{23}}{k_{13}} & -\dfrac{f_3}{k_{13}} \\ 0 & 1+\dfrac{k_{23}}{k_{12}}+\dfrac{k_{23}}{k_{13}} & -1-\dfrac{k_{23}}{k_{12}}-\dfrac{k_{23}}{k_{13}} & \dfrac{f_3}{k_{13}}-\dfrac{f_2}{k_{12}} \\ 0 & 0 & 0 & f_1+f_2+f_3 \end{bmatrix}, \tag{36}$$

which result reveals two possibilities.

(i) If $f_1+f_2+f_3 \neq 0$, then there is *no* solution. That mathematical result makes perfect sense physically, because with the end springs removed $f_1+f_2+f_3$ is the net lateral force on the three-mass system, and if that net force is nonzero, then the system cannot be in static equilibrium, as was assumed when we set $x_1'' = x_2'' = x_3'' = 0$ in (31)!

(ii) If $f_1+f_2+f_3 = 0$, then we see from (36) that there is an *infinity* of solutions, of the form

$$x_3 = \alpha, \qquad x_2 = \alpha + \text{etc.}, \qquad x_1 = \alpha + \text{etc.},$$

where α is an arbitrary constant and the two etc.'s involve the f's and k's. That is, the solution is nonunique because of the arbitrary *translation* α. Again, that result makes sense physically because with $k_1 = k_3 = 0$ there are no end springs to restrain the three-mass system laterally.

Let us make one more important point. Observe from (32) that the $\mathbf{K}$ matrix is symmetric. Yet the system (Fig. 1) is not *physically* symmetric; that is, in general, $k_1 \neq k_3$ and $k_{12} \neq k_{23}$. Thus, the mathematical symmetry is somewhat unexpected and mysterious. In fact, we state without proof that for *any* number of masses interconnected with springs the resulting $\mathbf{K}$ matrix will be symmetric.

There is a striking consequence of the symmetry of $\mathbf{K}$, which we now explain. Property I2 gives $(\mathbf{K}^{-1})^{\mathrm{T}} = (\mathbf{K}^{\mathrm{T}})^{-1} = \mathbf{K}^{-1}$ since $\mathbf{K}^{\mathrm{T}} = \mathbf{K}$. Let us denote

$\mathbf{K} = \{\alpha_{ij}\}$, say, and let us compare the displacement x_3 of m_3 due to a unit load $f_1 = 1$ on m_1 (with $f_2 = f_3 = 0$) with the displacement x_1 of m_1 due to a unit load $f_3 = 1$ on m_3 (with $f_1 = f_2 = 0$). In the first case,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

gives $x_3 = \alpha_{31}$ and, in the second case,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

gives $x_1 = \alpha_{13}$. But these are the same ($\alpha_{31} = \alpha_{13}$) because $\mathbf{K}$ is symmetric. In this manner we find that

$$\begin{pmatrix} \text{displacement } x_j \text{ of mass } m_j \\ \text{due to unit load on mass } m_k \end{pmatrix} = \begin{pmatrix} \text{displacement } x_k \text{ of mass } m_k \\ \text{due to unit load on mass } m_j \end{pmatrix}. \tag{37}$$

The latter "reciprocity" result can be generalized so as to apply to any linear elastic system and is known as *Maxwell reciprocity*.*

There is an electrical analog of the mechanical system shown in Fig. 1, a circuit containing resistors and voltage sources (such as batteries), and discussion of that case is left for the exercises.

10.6.3. Cramer's rule. We have seen that if $\mathbf{A}$ is $n \times n$ and det$\mathbf{A} \neq 0$, then $\mathbf{Ax} = \mathbf{c}$ has the unique solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}. \tag{38}$$

To focus on the individual components of $\mathbf{x}$, rather than the entire $\mathbf{x}$ vector, let us write out (38):

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \sum_j \alpha_{1j} c_j \\ \vdots \\ \sum_j \alpha_{nj} c_j \end{bmatrix}. \tag{39}$$

Equating the ith component on the left with the ith component on the right, we have the scalar statement

$$x_i = \sum_j \alpha_{ij} c_j \tag{40}$$

*By a linear elastic system is meant one obeying Hooke's law. For a statement and proof of Maxwell's reciprocity theorem see, for instance, Den Hartog's *Advanced Strength of Materials* (New York: McGraw-Hill, 1952).

for any desired i $(1 \leq i \leq n)$. Or, recalling (17),

$$x_i = \sum_j \left(\frac{A_{ji}}{\det \mathbf{A}} \right) c_j = \frac{\sum_j A_{ji} c_j}{\det \mathbf{A}}. \tag{41}$$

Now, if the numerator on the right-hand side were $\sum_j A_{ji} a_{ji}$, instead, it would be recognizable as the determinant of $\mathbf{A}$, namely, the cofactor expansion about the ith column. But the a_{ji}'s are replaced, in (41), by the c_j's, so the numerator of (41) amounts to a determinant but not the determinant of $\mathbf{A}$; rather, it is the determinant of the $\mathbf{A}$ matrix with its ith column replaced by the column of c_j's (or the $\mathbf{c}$ vector, if you like).

The result, known as **Cramer's rule**, after *Gabriel Cramer* (1704–1752), is as follows.

THEOREM 10.6.3 *Cramer's Rule*
If $\mathbf{Ax} = \mathbf{c}$ where $\mathbf{A}$ is invertible, then each component x_i of $\mathbf{x}$ may be computed as the ratio of two determinants; the denominator is det$\mathbf{A}$, and the numerator is also the determinant of the $\mathbf{A}$ matrix but with the ith column replaced by $\mathbf{c}$.

EXAMPLE 3. Let us solve the system

$$\begin{bmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \\ -2 \end{bmatrix} \tag{42}$$

for x_1 and x_2, say, using Cramer's rule. In this case det$\mathbf{A} = 8 \neq 0$ so the method is, first of all, applicable. Thus,

$$x_1 = \frac{\begin{vmatrix} \mathbf{5} & 3 & 0 \\ \mathbf{1} & 3 & 1 \\ \mathbf{-2} & 1 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{vmatrix}} = \frac{1}{8} \quad \text{and} \quad x_2 = \frac{\begin{vmatrix} 1 & \mathbf{5} & 0 \\ -2 & \mathbf{1} & 1 \\ 0 & \mathbf{-2} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{vmatrix}} = \frac{13}{8}, \tag{43}$$

where we have printed the "replacement columns" as boldface, for emphasis. ▮

Cramer's rule, like the inverse matrix solution (38) from which it comes, has the advantage of being an explicit *formula*, rather than a *method*. It is also useful in that it permits us to focus on any single component of $\mathbf{x}$ without having to compute the entire $\mathbf{x}$ vector.

10.6.4. Evaluation of $\mathbf{A}^{-1}$ by elementary row operations. Equation (16) gives $\mathbf{A}^{-1}$ in terms of det$\mathbf{A}$ and n^2 cofactors, each of which is ±1 times an $(n-1) \times$

$(n-1)$ minor determinant. Each of these determinants can be evaluated by the cofactor expansion definition or, especially if n is large, by a faster method – such as triangularization. Alternatively, we can bypass (16) altogether, and determine $\mathbf{A}^{-1}$ efficiently as follows.

Whether we are seeking $\mathbf{A}^{-1}$ in order to solve a system $\mathbf{Ax} = \mathbf{c}$, or whether we are simply seeking the inverse of a given matrix $\mathbf{A}$, observe that if we solve a system $\mathbf{Ax} = \mathbf{c}$ of n equations in n unknowns, or equivalently $\mathbf{Ax} = \mathbf{Ic}$, by Gauss–Jordan reduction, the result is the form $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$, or equivalently $\mathbf{Ix} = \mathbf{A}^{-1}\mathbf{c}$. Symbolically, then, the sequence of elementary row operations effects the following transformation:

$$\begin{gathered}\mathbf{Ax} = \mathbf{Ic} \\ \downarrow \\ \mathbf{Ix} = \mathbf{A}^{-1}\mathbf{c}.\end{gathered} \tag{44}$$

That is, at the same time that the row operations are transforming $\mathbf{A}$ to $\mathbf{I}$ they are also transforming $\mathbf{I}$ to $\mathbf{A}^{-1}$. Thus, we can skip $\mathbf{x}$ and $\mathbf{c}$ altogether, put $\mathbf{A}$ and $\mathbf{I}$ "side by side" as an augmented matrix $\mathbf{A}|\mathbf{I}$, and carry out elementary row operations on $\mathbf{A}|\mathbf{I}$ so as to reduce $\mathbf{A}$, on the left, to $\mathbf{I}$. When that has been accomplished, the matrix on the right will be $\mathbf{A}^{-1}$.

EXAMPLE 4. To illustrate, let us find the inverse of

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \tag{45}$$

Then

$$\begin{aligned}
\mathbf{A}|\mathbf{I} &= \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ -2 & 3 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \\
&\rightarrow \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 0 & \frac{8}{9} & -\frac{2}{9} & -\frac{1}{9} & 1 \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \\
&\rightarrow \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 0 & \frac{9}{4} & \frac{9}{8} & -\frac{9}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \rightarrow \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \\
&\rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{4} & -\frac{3}{8} & \frac{3}{8} \\ 0 & 1 & 0 & \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right]
\end{aligned} \tag{46}$$

so

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{4} & -\frac{3}{8} & \frac{3}{8} \\ \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{bmatrix}. \quad \blacksquare \tag{47}$$

Can this method *fail* to work? Yes indeed, it had *better* fail if det$\mathbf{A} = 0$, for then $\mathbf{A}$ is not invertible. The way that circumstance would show up is that the elementary row operations would produce one or more rows of zeros on the left so that $\mathbf{A}$ cannot be converted to $\mathbf{I}$.

10.6.5. LU-factorization. This final subsection is not really about the inverse matrix or about the inverse matrix method of solving $\mathbf{Ax} = \mathbf{c}$. Rather, it is about an alternative method of solution that is based upon the factorization of an $n \times n$ matrix $\mathbf{A}$ as a lower triangular matrix $\mathbf{L}$ times an upper triangular matrix $\mathbf{U}$:

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}, \tag{48}$$

where we have taken $n = 3$ simply for compactness. If we carry out the multiplication on the right and equate the nine elements of LU to the corresponding elements of $\mathbf{A}$ we obtain nine equations in the 12 unknown l_{ij}'s and u_{ij}'s. Since we have more unknowns than equations, there is some flexibility in implementing the idea. Hence there are various versions of LU-factorization.

According to **Doolittle's method** we can set each $l_{jj} = 1$ in $\mathbf{L}$ (i.e., the diagonal elements) and solve uniquely for the remaining l_{ij}'s and the u_{ij}'s. With $\mathbf{L}$ and $\mathbf{U}$ determined, we then solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{c}$ by setting $\mathbf{Ux} = \mathbf{y}$ so that $\mathbf{L}(\mathbf{Ux}) = \mathbf{c}$ breaks into the two problems

$$\mathbf{Ly} = \mathbf{c}, \tag{49a}$$

$$\mathbf{Ux} = \mathbf{y}, \tag{49b}$$

each of which is simple because $\mathbf{L}$ and $\mathbf{U}$ are triangular. We solve (49a) for $\mathbf{y}$, put that $\mathbf{y}$ into (49b), and then solve (49b) for $\mathbf{x}$. Let us illustrate the procedure.

EXAMPLE 5. To solve

$$\begin{bmatrix} 2 & -3 & 3 \\ 6 & -8 & 7 \\ -2 & 6 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix} \tag{50}$$

by the Doolittle LU-factorization method, we first need to determine $\mathbf{L}$ and $\mathbf{U}$ by equating

$$\begin{aligned} \begin{bmatrix} 2 & -3 & 3 \\ 6 & -8 & 7 \\ -2 & 6 & -1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \\ &= \begin{bmatrix} \underline{u_{11}} & \underline{u_{12}} & \underline{u_{13}} \\ \underline{l_{21}}u_{11} & l_{21}u_{12} + \underline{u_{22}} & l_{21}u_{13} + \underline{u_{23}} \\ \underline{l_{31}}u_{11} & l_{31}u_{12} + \underline{l_{32}}u_{22} & l_{31}u_{13} + l_{32}u_{23} + \underline{u_{33}} \end{bmatrix}. \end{aligned} \tag{51}$$

Matching $a_{11}, a_{12}, a_{13}, a_{21}, \ldots, a_{32}, a_{33}$ to the corresponding terms on the right gives a sequence of equations for $u_{11}, u_{12}, u_{13}, l_{21}, u_{22}, u_{23}, l_{31}, l_{32}$, and u_{33} (i.e., the underlined entries) in turn:

$$\begin{aligned}
u_{11} &= 2,\\
u_{12} &= -3,\\
u_{13} &= 3,\\
l_{21} &= 6/u_{11} = 6/2 = 3,\\
u_{22} &= -8 - l_{21}u_{12} = -8 - (3)(-3) = 1,\\
u_{23} &= 7 - l_{21}u_{13} = 7 - (3)(3) = -2,\\
l_{31} &= -2/u_{11} = -2/2 = -1,\\
l_{32} &= (6 - l_{31}u_{12})/u_{22} = [6 - (-1)(-3)]/1 = 3,\\
u_{33} &= -1 - l_{31}u_{13} - l_{32}u_{23} = -1 - (-1)(3) - (3)(-2) = 8.
\end{aligned} \tag{52}$$

Then (49a) becomes

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix},$$

which gives $\mathbf{y} = [-2, 3, -8]^{\mathrm{T}}$. Finally, (49a) becomes

$$\begin{bmatrix} 2 & -3 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix},$$

which gives the final solution $\mathbf{x} = [2, 1, -1]^{\mathrm{T}}$. ∎

The beauty of the method is that the l_{ij}'s and u_{ij}'s are found not by solving simultaneous coupled equations but by solving a sequence of linear equations in only one unknown [as illustrated in (52)]. With **L** and **U** thus determined, the solution of (49a) and (49b) is likewise simple since **L** and **U** are triangular. In fact, the method is around twice as fast as Gauss–Jordan elimination.

Closure. The inverse of a matrix **A**, denoted as $\mathbf{A}^{-1}$, exists if and only if **A** is square ($n \times n$) and det**A** $\neq 0$. If it exists it is given uniquely as

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\,\mathrm{adj}\mathbf{A}, \tag{53}$$

where the matrix adj**A** is the adjoint of **A** (the transpose of the cofactor matrix), and is such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{54}$$

The case where **A** is not invertible (i.e., is singular) is the exceptional case; in the generic case a given $n \times n$ matrix is invertible. Besides (54), several useful properties of inverses are given as I1–I4 in Section 10.6.1.

If **A** is invertible, then $\mathbf{Ax} = \mathbf{c}$ admits the unique solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} = \frac{1}{\det\mathbf{A}}\,(\mathrm{adj}\mathbf{A})\,\mathbf{c}, \tag{55}$$

which result gives us Cramer's rule, whereby each component of $\mathbf{x}$ is expressed as the ratio of $n \times n$ determinants.

Notice that the equation $ax = c$ has a unique solution if and only if $a \neq 0$. For $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is $n \times n$, that condition generalizes not to $\mathbf{A} \neq \mathbf{0}$ but to $\det\mathbf{A} \neq 0$.

In contrast with Gauss–Jordan reduction and LU-factorization, which are solution *methods*, (55) and Cramer's rule are explicit *formulas* for the solution (when a unique solution does exist).

Finally, we urge you to be careful with the sequencing of matrices because of the general absence of commutativity under multiplication. For instance, $\mathbf{Ax} = \mathbf{c}$ implies $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ (if $\mathbf{A}$ is invertible), NOT $\mathbf{x} = \mathbf{c}\mathbf{A}^{-1}$. Indeed, the product $\mathbf{c}\mathbf{A}^{-1}$ is not even defined (unless $n = 1$) since $\mathbf{c}$ is $n \times 1$ and $\mathbf{A}^{-1}$ is $n \times n$.

Computer software. Using *Maple*, the relevant command is **inverse(A)**, within the linalg package. For instance, to invert

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

enter

```
with(linalg):
```

and return. Then the steps

```
A := array ([[1, 2], [3, 4]]):
```

and

```
inverse(A);
```

give the result

$$\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

EXERCISES 10.6

1. Use (17) to evaluate the inverse matrix. If the matrix is not invertible, state that.

(a) $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $ad - bc \neq 0$

(b) $\begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix}$

(e) $\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$

(f) $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 0 & 3 & -1 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 1 & 0 \\ 2 & 0 & 5 \\ 0 & 0 & 3 \end{bmatrix}$

(h) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

(i) $\begin{bmatrix} 0 & 0 & 2 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

(j) $\begin{bmatrix} 3 & 1 & -2 \\ 1 & 2 & 1 \\ 1 & -3 & -3 \end{bmatrix}$

(k) $\begin{bmatrix} 7 & 1 & 3 \\ 2 & -1 & 1 \\ 0 & 1 & 4 \end{bmatrix}$

(l) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 4 & 0 & 5 \end{bmatrix}$ (m) $\begin{bmatrix} 2 & 3 & -5 \\ 12 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix}$

(n) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix}$ (o) $\begin{bmatrix} 2 & 3 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix}$

(p) $\begin{bmatrix} 7 & 1 & 3 & 0 \\ 2 & -1 & 1 & 0 \\ 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (q) $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

(r) $\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (s) $\begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$

2. (c) – (o) Evaluate the inverse of the matrix given in the corresponding part of Exercise 1 using elementary row operations, as we did in Example 4.

3. (c) – (o) Evaluate the inverse of the matrix given in the corresponding part of Exercise 1 using computer software.

4. (*Block-diagonal matrices*) If an $n \times n$ matrix $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & \cdots & \mathbf{A}_k \end{bmatrix} \qquad (k \leq n)$$

it is said to be **block diagonal**. All of the $\mathbf{A}_j$ submatrices need to be square, although not necessarily of equal order, with their main diagonals coinciding with the main diagonal of $\mathbf{A}$. For instance,

$$\mathbf{A} = \left[\begin{array}{cc|ccc|c} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 3 & 0 \\ 0 & 0 & 2 & 5 & 1 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 4 \end{array}\right] = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix} \tag{4.1}$$

is block diagonal. Such matrices exhibit essentially the same simple features as diagonal matrices.

(a) Show that $\mathbf{A}$ is invertible if and only if $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are. HINT: Recall equation (9.1) in Exercise 9, Section 10.4.

(b) Assuming that $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are invertible, verify that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_1^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2^{-1} & & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & \cdots & \mathbf{A}_k^{-1} \end{bmatrix}.$$

(c) Use the latter result to evaluate $\mathbf{A}^{-1}$, where $\mathbf{A}$ is given in (4.1).

5. Solve for x_1 and x_2 by Cramer's rule.

(a) $$\begin{aligned} x_1 + 4x_2 &= 0 \\ 3x_1 - x_2 &= 6 \end{aligned}$$

(b) $$\begin{aligned} ax_1 + bx_2 &= c \\ dx_1 + ex_2 &= f \end{aligned}$$

(c) $$\begin{aligned} x_1 - 2x_2 + x_3 &= 4 \\ 2x_1 + 3x_2 + x_3 &= -7 \\ 4x_1 + x_2 + 2x_3 &= 0 \end{aligned}$$

(d) $$\begin{aligned} x_1 + 2x_2 + 3x_3 &= 9 \\ x_1 + 4x_2 &= 6 \\ x_1 - 5x_3 &= 2 \end{aligned}$$

(e) $$\begin{aligned} 2x_1 + x_2 &= 1 \\ x_1 + 2x_2 + x_3 &= 0 \\ x_2 + 2x_3 + x_4 &= 0 \\ x_3 + 2x_4 &= 0 \end{aligned}$$

(f) $$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + 2x_2 + 3x_3 &= 0 \\ x_1 - x_2 + 4x_3 &= 0 \end{aligned}$$

6. (a) Given a certain 3×3 matrix $\mathbf{A}$, we find its inverse to be

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

Can that result be correct? Explain.
(b) Same as (a), for

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 0 & 0 \end{bmatrix}.$$

7. If $\mathbf{A}^{-1}$ is the given matrix, find $\mathbf{A}$.

(a) $\begin{bmatrix} 3 & -1 \\ 3 & -2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 1 \\ 6 & 2 & 1 \end{bmatrix}$

8. Suppose that $\mathbf{Ax} = \mathbf{c}$ is a linear system of order 3, and that to the $\mathbf{c}$ vectors

$$\mathbf{c} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

there correspond unique solutions

$$\mathbf{x} = \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix},$$

respectively. Then what is the solution $\mathbf{x}$ corresponding to $\mathbf{c} = [4, 3, -1]^{\mathrm{T}}$? Is it unique? Explain.

9. Suppose that Gauss elimination gives the solution of a linear system $\mathbf{Ax} = \mathbf{c}$ as $\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2$, where $\mathbf{A}$ is 6×6 and α_1 and α_2 are arbitrary. Is $\mathbf{A}$ invertible? Explain.

10. (*Nilpotent matrices*) If there is some positive integer p such that $\mathbf{A}^p = \mathbf{0}$, then $\mathbf{A}$ said to be **nilpotent** (i.e., potentially nil).

(a) Show that a nilpotent matrix is necessarily singular.
(b) If $\mathbf{A}$ is nilpotent, with $\mathbf{A}^p = \mathbf{0}$, show that

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{p-1}. \qquad (10.1)$$

11. First, read Exercise 10. Use (10.1) to find the inverse of the given matrix. HINT: You will need to identify $\mathbf{A}$.

(a) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 8 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 2 & 7 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 5 & 1 & 0 \\ 0 & 1 & 2 & 7 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ -4 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 2 & 1 & 6 & 1 \end{bmatrix}$

12. First, read Exercise 10. Use (10.1) to find the inverse of the given matrix

(a) $\begin{bmatrix} 2 & 8 & 10 \\ 0 & 3 & 12 \\ 0 & 0 & 4 \end{bmatrix}$ HINT:

$$\begin{bmatrix} 2 & 8 & 10 \\ 0 & 3 & 12 \\ 0 & 0 & 4 \end{bmatrix}^{-1} = \left[\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}\begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}\right]^{-1}$$
$$= \begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

(b) $\begin{bmatrix} 3 & 0 & 0 \\ 4 & -2 & 0 \\ 10 & 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 0 & 0 \\ 4 & 2 & 0 \\ 1 & 6 & 2 \end{bmatrix}$

13. (*Ill-conditioned systems*) Consider the system

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad \text{or} \quad \mathbf{Ax} = \mathbf{c}. \quad (13.1)$$

Here $\mathbf{A}$ is a third-order **Hilbert matrix**, named after *David Hilbert* (1862–1943).

(a) Evaluate $\mathbf{A}^{-1}$, by any analytical means that you wish, and show that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} = [-3, -12, 30]^{\mathrm{T}}$.
(b) To simulate the effects of roundoff error, consider in place of (13.1) the rounded off system

$$\begin{bmatrix} 1 & 0.5 & 0.33 \\ 0.5 & 0.33 & 0.25 \\ 0.33 & 0.25 & 0.2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}. \qquad (13.2)$$

Solving (13.2) by any means you wish (computer software being the easiest), show that the solution of (13.2) is $\mathbf{x} \approx [11.1, -84.1, 96.8]^{\mathrm{T}}$. NOTE: In this example we see that only a slight error in $\mathbf{A}$ leads to a disproportionately large error in the solution. Hence, (13.1) is said to be **ill-conditioned**. (Ill-conditioned systems are also mentioned in Exercise 13, Section 8.3.) In applications it is important to know if a given system is ill-conditioned so that steps can be taken to obtain a sufficiently accurate solution. According to one criterion in the literature, an $n \times n$ matrix $\mathbf{A}$ may be considered as ill-conditioned if

$$\frac{|\det\mathbf{A}|}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}} \ll 1. \qquad (13.3)$$

For the Hilbert matrix in (13.1), the left-hand side of (13.3) is 0.00033, which is indeed much smaller than unity.
(c) In place of (13.2), use the more accurate rounded off system

$$\begin{bmatrix} 1 & 0.5 & 0.333 \\ 0.5 & 0.333 & 0.25 \\ 0.333 & 0.25 & 0.2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \qquad (13.4)$$

and see how much closer the solution of (13.4) comes to the exact solution $\mathbf{x} = [-3, -12, 30]^{\mathrm{T}}$ of (13.1).

14. Prove property I3.

15. Prove property I4.

16. In Section 10.6.2 we stated that if we are willing to remove two springs then we can have det$\mathbf{K} = 0$ either by setting $k_1 = k_3 = 0$ or by setting $k_{12} = k_{23} = 0$. We discuss the former choice, $k_1 = k_3 = 0$, both mathematically and physically as well. Do the same for the latter choice, $k_{12} = k_{23} = 0$.

17. (*A dc circuit*) Application of Kirchhoff's laws to the circuit shown in Exercise 10 of Section 8.3 produced the five equations (10.1) on the currents i_1, i_2, i_3. Of these equations, the second is the same as the first, and the fifth is the fourth minus the third. Thus, deleting those two redundant equations leaves the system

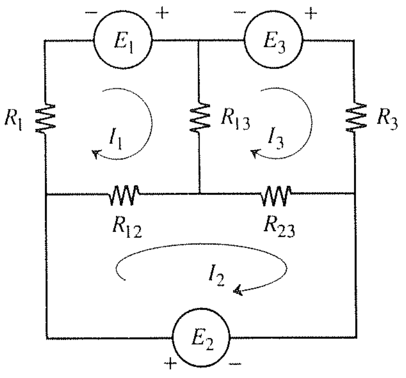
$$\begin{aligned} i_1 - \quad i_2 - \quad i_3 &= 0 \\ R_2 i_2 - R_3 i_3 &= 0 \\ R_1 i_1 + R_2 i_2 \qquad &= E. \end{aligned} \tag{17.1}$$

(a) Show that the determinant of the coefficient matrix in (17.1) is necessarily nonzero, so that the system $\mathbf{Ri} = \mathbf{e}$ given by (17.1) necessarily admits the unique solution $\mathbf{i} = \mathbf{R}^{-1}\mathbf{e}$. NOTE: $R_1 > 0, R_2 > 0$, and $R_3 > 0$.
(b) Solve for $\mathbf{i}$ by the inverse matrix method. Also, solve for i_1, i_2, i_3 by Cramer's rule and verify that the results are the same.
(c) Suppose, instead, that we allow one or more of the resistances to be zero so that $R_1 \geq 0, R_2 \geq 0, R_3 \geq 0$. Show that if any two, or all three, of the resistances are zero, then the determinant *does* vanish so that equations (17.1) admit either no solution or an infinity of solutions. For each of these four "singular" cases determine whether there is no solution or an infinity of solutions by applying Gauss elimination. State the *physical* significance of each of these results insofar as possible.

18. (*Circuit analog*) The electrical circuit analog of the mass-spring system shown in Fig. 1 is shown below.
(a) Applying Kirchhoff's voltage law, show that

$$\begin{bmatrix} (R_1 + R_{12} + R_{13}) & -R_{12} & -R_{13} \\ -R_{12} & (R_{12} + R_{23}) & -R_{23} \\ -R_{13} & -R_{23} & (R_{13} + R_{23} + R_3) \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}, \tag{18.1}$$

which is the analog of (32) under the correspondence $R_{ij} \leftrightarrow k_{ij}, I_j \leftrightarrow x_j, E_j \leftrightarrow f_j$.
(b) Discuss the existence and uniqueness of solutions of (18.1) in the same way that we did that for the mass-spring system in Section 10.6.2, including a reciprocity result analogous to (37).



19. Solve by the Doolittle LU-factorization method.

(a) $\begin{bmatrix} 2 & 3 \\ 8 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -4 \\ 10 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 13 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 5 & 1 \\ 2 & 8 & 0 \\ 8 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -7 \\ 10 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & 3 & -1 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ 12 \end{bmatrix}$

20. (*Dual or reciprocal set*) In Exercise 13 of Section 9.9 we introduced the concept of a set of dual or reciprocal vectors $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*\}$ corresponding to a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ that is not necessarily orthogonal, or ON. Having learned more about the solution of systems of linear algebraic equations in Sections 10.5 and 10.6, we can now return to that exercise and prove the claim made in part (a). Specifically, prove the dual set exists, is unique, and is itself a basis for $\mathbb{R}^n$.

10.7 Change of Basis (Optional)

In a given problem one selects what appears to be the most convenient basis, but at some stage of the analysis it may be desirable to switch to some other basis. For example, in studying the aerodynamics of a propeller it is generally most convenient to carry out the analysis (of the propeller-induced pressure field, for example) with respect to a propeller-fixed basis, one that rotates with the propeller, although eventually we may wish to relate quantities back to a stationary (nonrotating) basis. How do the coordinates (i.e., the components) of a given vector change as we change the basis? It is that question which we address in this section.

Let $B = \{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ be a given basis for the vector space V under consideration so that any given vector $\mathbf{x}$ in V can be expanded as

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n. \tag{1}$$

If we switch to some other basis $B' = \{\mathbf{e}'_1, \dots, \mathbf{e}'_n\}$, then we may, similarly, expand the same vector $\mathbf{x}$ as

$$\mathbf{x} = x'_1\mathbf{e}'_1 + \cdots + x'_n\mathbf{e}'_n. \tag{2}$$

How are the x'_j coordinates related to the x_j coordinates? Since B' is a basis, we may expand each of the $\mathbf{e}_j$'s in terms of B':

$$\begin{aligned} \mathbf{e}_1 &= q_{11}\mathbf{e}'_1 + \cdots + q_{n1}\mathbf{e}'_n, \\ &\vdots \\ \mathbf{e}_n &= q_{1n}\mathbf{e}'_1 + \cdots + q_{nn}\mathbf{e}'_n. \end{aligned} \tag{3}$$

Putting (3) into (1) gives

$$\begin{aligned} \mathbf{x} &= x_1\left(q_{11}\mathbf{e}'_1 + \cdots + q_{n1}\mathbf{e}'_n\right) + \cdots + x_n\left(q_{1n}\mathbf{e}'_1 + \cdots + q_{nn}\mathbf{e}'_n\right) \\ &= \left(x_1q_{11} + \cdots + x_nq_{1n}\right)\mathbf{e}'_1 + \cdots + \left(x_1q_{n1} + \cdots + x_nq_{nn}\right)\mathbf{e}'_n, \end{aligned} \tag{4}$$

and a comparison of (2) and (4) gives the desired relations

$$\begin{aligned} x'_1 &= q_{11}x_1 + \cdots + q_{1n}x_n, \\ &\vdots \\ x'_n &= q_{n1}x_1 + \cdots + q_{nn}x_n \end{aligned} \tag{5}$$

or, in matrix notation,

$$\boxed{[\mathbf{x}]_{B'} = \mathbf{Q}\,[\mathbf{x}]_B\,,} \tag{6}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \tag{7}$$

and

$$[\mathbf{x}]_B = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad [\mathbf{x}]_{B'} = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}. \tag{8}$$

We call $[\mathbf{x}]_B$ the **coordinate vector** of the vector $\mathbf{x}$ with respect to the ordered basis B, and similarly for $[\mathbf{x}]_{B'}$, and we call $\mathbf{Q}$ the **coordinate transformation matrix** from B to B'.

Thus far, our results apply whether the bases are orthogonal or not. *In the remainder of this section we assume that both bases, B and B', are ON.* Thus, let us rewrite (3) as

$$\begin{aligned} \hat{\mathbf{e}}_1 &= q_{11}\hat{\mathbf{e}}_1' + \cdots + q_{n1}\hat{\mathbf{e}}_n', \\ &\vdots \\ \hat{\mathbf{e}}_n &= q_{1n}\hat{\mathbf{e}}_1' + \cdots + q_{nn}\hat{\mathbf{e}}_n', \end{aligned} \tag{9}$$

where the carets denote unit vectors, as usual. If we dot $\hat{\mathbf{e}}_1'$ into both sides of the first equation in (9), and remember that B' is ON, we obtain $q_{11} = \hat{\mathbf{e}}_1' \cdot \hat{\mathbf{e}}_1$. Dotting $\hat{\mathbf{e}}_2'$ gives $q_{21} = \hat{\mathbf{e}}_2' \cdot \hat{\mathbf{e}}_1, \dots,$ dotting $\hat{\mathbf{e}}_n'$ gives $q_{n1} = \hat{\mathbf{e}}_n' \cdot \hat{\mathbf{e}}_1$, and similarly for the second through nth equation in (9). The result is the formula

$$\boxed{q_{ij} = \hat{\mathbf{e}}_i' \cdot \hat{\mathbf{e}}_j,} \tag{10}$$

which tells us how to compute the transformation matrix $\mathbf{Q}$.

There are two properties of the $\mathbf{Q}$ matrix to address before turning to an example. To obtain the first of these, observe that

$$\begin{aligned} \mathbf{Q}^{\mathrm{T}}\mathbf{Q} &= \left[\begin{array}{ccc} q_{11} & \cdots & q_{n1} \\ \hline q_{12} & \cdots & q_{n2} \\ \hline & \vdots & \\ \hline q_{1n} & \cdots & q_{n2} \end{array}\right] \left[\begin{array}{c|c|c|c} q_{11} & q_{12} & & q_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1} & q_{n2} & & q_{nn} \end{array}\right] \\ &= \begin{bmatrix} \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_n \\ \vdots & & \vdots \\ \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_n \end{bmatrix} = \mathbf{I} \end{aligned} \tag{11}$$

so that

$$\boxed{\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}.} \tag{12}$$

It was useful to partition the $\mathbf{Q}^{\mathrm{T}}$ and $\mathbf{Q}$ matrices in (11) because we can see from (9) that the columns of $\mathbf{Q}$ (and hence the rows of $\mathbf{Q}^{\mathrm{T}}$) are actually the $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_n$ vectors (in n-tuple form). Thus, from the way matrix multiplication is defined, we can see that the elements of the product matrix are dot products. Specifically, the i, j element of $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ is $\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j$, which is the Kronecker delta δ_{ij}. Hence, $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ equals the identity matrix $\mathbf{I}$, so $\mathbf{Q}^{\mathrm{T}}$ must be the inverse of $\mathbf{Q}$, as stated in (12).

That result makes it easy for us to reverse equation (6) – that is, to solve for $[\mathbf{x}]_B$ in terms of $[\mathbf{x}]_{B'}$ for then $[\mathbf{x}]_B = \mathbf{Q}^{-1}[\mathbf{x}]_{B'} = \mathbf{Q}^{\mathrm{T}}[\mathbf{x}]_{B'}$. In other words, we do not need to face up to the evaluation of $\mathbf{Q}^{-1}$ since $\mathbf{Q}^{-1}$ is merely $\mathbf{Q}^{\mathrm{T}}$.

Any matrix with the useful property (12) is known as an **orthogonal matrix** because it follows from (12) that the column vectors in $\mathbf{Q}$ are orthonormal.

As the second property of $\mathbf{Q}$, observe that it also follows from (12) that

$$\boxed{\det\mathbf{Q} = \pm 1,} \tag{13}$$

that is, either $+1$ or -1 since $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ implies that $\det(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}) = \det\mathbf{I} = 1$. But $\det(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}) = (\det\mathbf{Q}^{\mathrm{T}})(\det\mathbf{Q}) = (\det\mathbf{Q})(\det\mathbf{Q}) = (\det\mathbf{Q})^2$. Hence, $\det\mathbf{Q}$ must be $+1$ or -1.

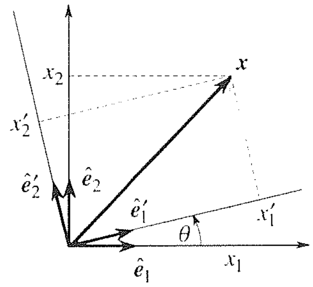


Figure 1. Rotation in the plane.

EXAMPLE 1. *Rotation in the Plane.* Consider the vector space $\mathbb{R}^2$, with the ON bases $B = \{\hat{\mathrm{e}}_1, \hat{\mathrm{e}}_2\}$ and $B' = \{\hat{\mathrm{e}}_1', \hat{\mathrm{e}}_2'\}$ shown in Fig. 1. B' is obtained from B by a counterclockwise rotation through an angle θ (or clockwise if θ is negative). From the figure,

$$\begin{aligned} q_{11} &= \hat{\mathrm{e}}_1' \cdot \hat{\mathrm{e}}_1 = (1)(1)\cos\theta = \cos\theta, \\ q_{12} &= \hat{\mathrm{e}}_1' \cdot \hat{\mathrm{e}}_2 = (1)(1)\cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\ q_{21} &= \hat{\mathrm{e}}_2' \cdot \hat{\mathrm{e}}_1 = (1)(1)\cos\left(\frac{\pi}{2} + \theta\right) = -\sin\theta, \\ q_{22} &= \hat{\mathrm{e}}_2' \cdot \hat{\mathrm{e}}_2 = (1)(1)\cos\theta = \cos\theta, \end{aligned} \tag{14}$$

so that the coordinate transformation matrix is

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{15}$$

Hence,

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{16}$$

Or, the other way around,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^{-1} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

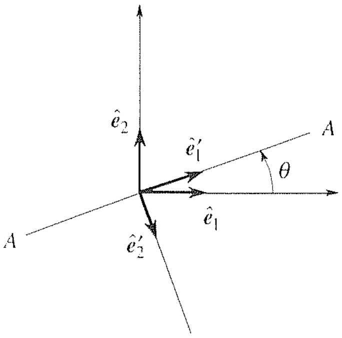


Figure 2. Rotation plus reflection.

so

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}. \tag{17}$$

COMMENT 1. It is easy enough to check (16) and (17) for one or two special cases. For example, if $\theta = 0$ then the two bases coincide so we should have $x_1' = x_1$ and $x_2' = x_2$, and that is what (16) and (17) give. Also, if $\theta = \pi/2$, say, we should have $x_1' = x_2$ and $x_2' = -x_1$ and, again, that is what (16) and (17) give.

COMMENT 2. Two ON bases in a plane are not necessarily related through a rotation. In this example, for instance, if we reverse the direction of $\hat{\mathrm{e}}_2'$, then $\{\hat{\mathrm{e}}_1', \hat{\mathrm{e}}_2'\}$ is still ON

(Fig. 2), but is not obtainable from $\{\hat{e}_1, \hat{e}_2\}$ by means of a rotation alone. Rather, we need a rotation *and a reflection*, a counterclockwise rotation through an angle θ, and then a reflection about AA (or, first a reflection about the $\hat{e}_1$ axis and then a counterclockwise rotation through an angle θ). In this case

$$\begin{aligned}
q_{11} &= \hat{e}'_1 \cdot \hat{e}_1 = \cos\theta, \\
q_{12} &= \hat{e}'_1 \cdot \hat{e}_2 = \cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\
q_{21} &= \hat{e}'_2 \cdot \hat{e}_1 = \cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\
q_{22} &= \hat{e}'_2 \cdot \hat{e}_2 = \cos(\pi - \theta) = -\cos\theta
\end{aligned}$$

so that $\mathbf{Q}$ is the orthogonal matrix

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}. \tag{18}$$

Recall from (13) that det$\mathbf{Q}$ is either $+1$ or -1. For the case where B and B' are related through a pure rotation [$\mathbf{Q}$ given by (15)] det$\mathbf{Q} = +1$, and for the case where they are related through a reflection and a rotation [$\mathbf{Q}$ given by (18)] det$\mathbf{Q} = -1$. ∎

Closure. In this brief section we study the relationship between the components, or coordinates, of any given vector $\mathbf{x}$ expanded in terms of two different bases B and B'. We find the linear relationship (6), where the q_{ij} elements of the coordinate transformation matrix $\mathbf{Q}$ are the expansion coefficients of $\mathbf{e}_j$ in terms of $\mathbf{e}'_1, \dots, \mathbf{e}'_n$, as indicated by (3).

If B and B' are ON, then the q_{ij}'s are computed, simply, from (10), and $\mathbf{Q}$ admits the properties that $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$ and that det$\mathbf{Q}$ is $+1$ or -1. *Any* matrix $\mathbf{Q}$ having the property $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$ has ON column vectors and is called an orthogonal matrix.

EXERCISES 10.7

1. Given that $\mathbf{e}_1 = \mathbf{e}'_1 + 2\mathbf{e}'_2$ and $\mathbf{e}_2 = \mathbf{e}'_1 - \mathbf{e}'_2$, find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal? If $[\mathbf{x}]_B = [5, -1]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$. If $[\mathbf{x}]_{B'} = [2, 3]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

2. Given that $\mathbf{e}_1 = \mathbf{e}'_1 + \mathbf{e}'_2 - \mathbf{e}'_3$, $\mathbf{e}_2 = \mathbf{e}'_1 - \mathbf{e}'_2 + \mathbf{e}'_3$, and $\mathbf{e}_3 = -\mathbf{e}'_1 + \mathbf{e}'_2 + \mathbf{e}'_3$, find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal? If $[\mathbf{x}]_B = [4, 1, -2]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$. If $[\mathbf{x}]_{B'} = [1, 0, 2]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

3. Let $\hat{e}_1 = [1, 0]^{\mathrm{T}}$, $\hat{e}_2 = [0, 1]^{\mathrm{T}}$, and $\hat{e}'_1 = \frac{1}{\sqrt{5}}[2, 1]^{\mathrm{T}}$, $\hat{e}'_2 = \frac{1}{\sqrt{5}}[1, -2]^{\mathrm{T}}$.

(a) Find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal?

(b) If $[\mathbf{x}]_B = [8, -6]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$.

(c) If $[\mathbf{x}]_{B'} = [1, 3]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

4. Let $\hat{e}_1 = [1, 0, 0, 0]^{\mathrm{T}}$, $\hat{e}_2 = [0, 1, 0, 0]^{\mathrm{T}}$, $\hat{e}_3 = [0, 0, 1, 0]^{\mathrm{T}}$, $\hat{e}_4 = [0, 0, 0, 1]^{\mathrm{T}}$, and $\hat{e}'_1 = \frac{1}{\sqrt{2}}[1, 1, 0, 0]^{\mathrm{T}}$, $\hat{e}'_2 = [0, 0, 1, 0]^{\mathrm{T}}$, $\hat{e}'_3 = \frac{1}{\sqrt{3}}[1, -1, 0, 1]^{\mathrm{T}}$, $\hat{e}'_4 = \frac{1}{\sqrt{6}}[1, -1, 0, -2]^{\mathrm{T}}$.

(a) Find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal?

(b) If $[\mathbf{x}]_B = [1, 1, 2, 5]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$.

(c) If $[\mathbf{x}]_{B'} = [1, 1, 2, 5]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

5. Show whether or not these matrices are orthogonal.

(a) $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ (f) $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} \end{bmatrix}$

6. For the case of rotation in a plane, the transformation $\mathbf{Q}$ corresponded to a counterclockwise rotation of the basis through an angle θ. Does $\mathbf{Q}^{-1}$ correspond to the reverse of this, a *clockwise* rotation θ ? Prove or disprove.

7. (*Rotation and reflection*) (a) Show that *every* orthogonal coordinate transformation matrix of order 2 is of one of the following two types:

$$\mathbf{Q}_1 = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad \mathbf{Q}_2 = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix},$$

i.e., as given by the pure rotation (15) or by the rotation plus reflection (18).

(b) Show that these two cases can be distinguished by the sign of the determinant, specifically, that det$\mathbf{Q} = +1$ if $\mathbf{Q}$ corresponds to pure rotation, and that det$\mathbf{Q} = -1$ if $\mathbf{Q}$ corresponds to rotation plus reflection.

8. (a) Prove that if $\mathbf{Q}$ is orthogonal, then so is $\mathbf{Q}^{\mathrm{T}}$.

(b) Prove that if $\mathbf{Q}$ is orthogonal, then so is $\mathbf{Q}^{-1}$.

9. Evaluate $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^n$.

10.8 Vector Transformation (Optional)

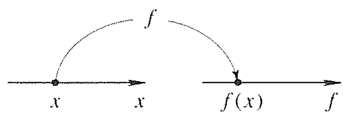


Figure 1. Function f as a transformation.

Recall that a real-valued function f of a real variable x is a rule that assigns a uniquely determined value $f(x)$ to each specified value x as illustrated in Fig. 1. Thus, f is a transformation, or mapping, from points on an x axis to points on an f axis, and we view x as the "input" and $f(x)$ as the "output." [A more familiar graphical display of f, called the *graph* of f, can be obtained if, following *Descartes* (1596–1650), we arrange the x and f axes at right angles to each other and plot the set of points $x, f(x)$ as illustrated in Fig. 2.]

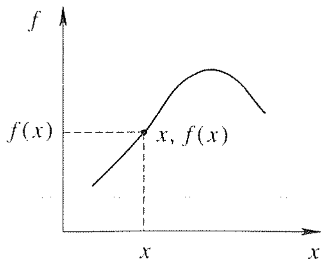


Figure 2. The graph of f.

In this section we reconsider vectors and matrices from this transformation point of view. Specifically, we consider vector-valued functions $\mathbf{F}$ of a vector variable $\mathbf{x}$. That is, the "input" is now a vector $\mathbf{x}$ from some vector space V, and the function $\mathbf{F}$ assigns a uniquely determined "output" vector $\mathbf{F}(\mathbf{x})$ in some vector space W. We call $\mathbf{F}$ a **transformation**, or **mapping**, from V into W, and denote it as

$$\mathbf{F}: V \to W.$$

We call V the **domain of F** and W the **range of definition of F**. W may, but need not, be identical to V. If it *is* identical, then $\mathbf{F} : V \to V$ is called an **operator** on V.

EXAMPLE 1. To illustrate, consider the transformation $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^3$ defined by

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix}. \tag{1}$$

Here V is $\mathbb{R}^4$, W is $\mathbb{R}^3$, and the input vector $\mathbf{x}$ and the output vector $\mathbf{F}(\mathbf{x})$ are

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \text{ in } \mathbb{R}^4, \quad \text{and} \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix} \text{ in } \mathbb{R}^3. \tag{2}$$

For example, if $\mathbf{x} = [2, 3, 6, -1]^{\mathrm{T}}$, then $\mathbf{F}(\mathbf{x}) = [-5, 8, 21]^{\mathrm{T}}$. This transformation is not an operator since $W = \mathbb{R}^3$, whereas $V = \mathbb{R}^4$. ▮

We say that the vector $\mathbf{F}(\mathbf{x})$ in W is the **image** of the vector $\mathbf{x}$ in V under the transformation $\mathbf{F}$, and that $\mathbf{x}$ is the **inverse image** of $\mathbf{F}(\mathbf{x})$.

Since V and W are vector spaces, each must contain a zero vector. We will denote these zero vectors as $\mathbf{0}_V$ and $\mathbf{0}_W$, respectively. Finally, we define the image of V in W as the **range** R of $\mathbf{F}$, and we define the inverse image of $\mathbf{0}_W$ in V as the **nullspace** or **kernel** K of $\mathbf{F}$. That is, the kernel K is the part of V that maps to the zero vector $\mathbf{0}_W$ in W.

EXAMPLE 2. Let us find the range and kernel of the transformation $\mathbf{F}$ given in Example 1. First, the range. The range of $\mathbf{F}$ is the set of all vectors $\mathbf{c}$ in W for which the equation $\mathbf{F}(\mathbf{x}) = \mathbf{c}$ is consistent, that is, has at least one solution $\mathbf{x}$ in V. In the present case $\mathbf{F}(\mathbf{x}) = \mathbf{c}$ is

$$\begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{3}$$

or, in scalar form,

$$\begin{aligned} x_1 - 2x_2 \qquad\quad + x_4 &= c_1, \\ x_1 \qquad\quad + \;\; x_3 \qquad\;\; &= c_2, \\ x_1 + 2x_2 + 2x_3 - x_4 &= c_3. \end{aligned} \tag{4}$$

Applying elementary operations to (4), we obtain the equivalent system

$$\begin{aligned} x_1 - 2x_2 \qquad\quad + x_4 &= c_1, \\ 2x_2 + x_3 - x_4 &= c_2 - c_1, \\ 0 &= c_3 - 2c_2 + c_1. \end{aligned} \tag{5a,b,c}$$

This system is consistent if and only if $\mathbf{c}$ lies in the plane (through the origin) defined by $c_3 - 2c_2 + c_1 = 0$. That plane is a two-dimensional subspace of $\mathbb{R}^3$ and is the range R of $\mathbf{F}$.

Turning to the kernel K of $\mathbf{F}$, that is, the inverse image of $[0,0,0]^{\mathrm{T}}$ in $\mathbb{R}^4$, we need merely set $c_1 = c_2 = c_3 = 0$ in (5) and solve for $\mathbf{x}$. That solution gives

$$\mathbf{x} = \alpha_1 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \end{bmatrix} + \alpha_2 \begin{bmatrix} -2 \\ -1 \\ 2 \\ 0 \end{bmatrix} \equiv \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2,$$

where α_1, α_2 are arbitrary constants. Then $K = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ and $\dim K = 2$. These results are summarized, schematically, in Fig. 3. ▮

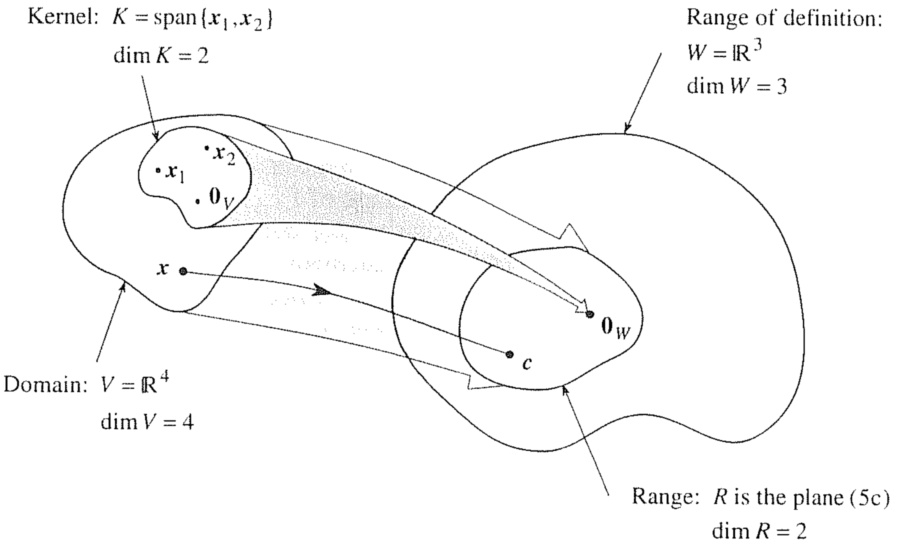


Figure 3. The transformation **F**.

Just as linearity, or the absence of it, was crucial in the theory of ordinary differential equations, it is likewise crucial here. We distinguish transformations as linear or nonlinear as follows.

DEFINITION 10.8.1 *Linear Transformation*
We say that $\mathbf{F} : V \to W$ is a **linear** transformation if

$$\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) \tag{6}$$

for every choice of vectors $\mathbf{u}, \mathbf{v}$ in V, and scalars α, β; otherwise, $\mathbf{F}$ is said to be **nonlinear**.

EXAMPLE 3. To illustrate, consider the transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$, where

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_3 \\ x_2 + 5x_3 \end{bmatrix}. \tag{7}$$

Let us see if $\mathbf{F}$ is linear.

$$\begin{aligned} \mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) &= \begin{bmatrix} (\alpha u_1 + \beta v_1) - 2(\alpha u_2 + \beta v_2) + (\alpha u_3 + \beta v_3) \\ (\alpha u_2 + \beta v_2) + 5(\alpha u_3 + \beta v_3) \end{bmatrix} \\ &= \alpha \begin{bmatrix} u_1 - 2u_2 + u_3 \\ u_2 + 5u_3 \end{bmatrix} + \beta \begin{bmatrix} v_1 - 2v_2 + v_3 \\ v_2 + 5v_3 \end{bmatrix} \\ &= \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) \end{aligned} \tag{8}$$

for every choice of $\mathbf{u}, \mathbf{v}, \alpha, \beta$, so $\mathbf{F}$ is indeed linear. Notice that the key step in (8), the second equality, follows from the definitions of the addition and scalar multiplication of matrices. ▮

EXAMPLE 4. Consider $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, where

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_1 + 2x_2 \end{bmatrix}. \tag{9}$$

Then

$$\begin{aligned} &\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) \\ &\quad = \begin{bmatrix} (\alpha u_1 + \beta v_1)^2 \\ (\alpha u_1 + \beta v_1) + 2(\alpha u_2 + \beta v_2) \end{bmatrix} \\ &\quad = \alpha \begin{bmatrix} u_1^2 \\ u_1 + 2u_2 \end{bmatrix} + \beta \begin{bmatrix} v_1^2 \\ v_1 + 2v_2 \end{bmatrix} + \begin{bmatrix} (\alpha^2 - \alpha)u_1^2 + (\beta^2 - \beta)v_1^2 + 2\alpha\beta u_1 v_1 \\ 0 \end{bmatrix} \\ &\quad = \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) + \text{deviation}, \end{aligned} \tag{10}$$

where the "deviation,"

$$\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) - \alpha\,\mathbf{F}(\mathbf{u}) - \beta\,\mathbf{F}(\mathbf{v}) = \begin{bmatrix} (\alpha^2 - \alpha)u_1^2 + (\beta^2 - \beta)v_1^2 + 2\alpha\beta u_1 v_1 \\ 0 \end{bmatrix} \tag{11}$$

is obviously not zero for all choices of $\alpha, \beta, \mathbf{u}, \mathbf{v}$. For instance, if $\alpha = \beta = u_1 = v_1 = 1$, then (for any u_2 and v_2) the deviation vector is $[2, 0]^{\mathrm{T}}$. Thus, (6) does not hold for all choices of $\alpha, \beta, \mathbf{u}, \mathbf{v}$, so $\mathbf{F}$ is nonlinear. ▮

If $\mathbf{F}$ is linear, then besides (6) we have

$$\begin{aligned} \mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v} + \gamma\mathbf{w}) &= \mathbf{F}(\alpha\mathbf{u} + (\beta\mathbf{v} + \gamma\mathbf{w})) \\ &= \alpha\,\mathbf{F}(\mathbf{u}) + \mathbf{F}(\beta\mathbf{v} + \gamma\mathbf{w}) = \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) + \gamma\,\mathbf{F}(\mathbf{w}) \end{aligned}$$

or, more generally,

$$\boxed{\mathbf{F}(\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n) = \alpha_1\mathbf{F}(\mathbf{u}_1) + \cdots + \alpha_n\mathbf{F}(\mathbf{u}_n).} \tag{12}$$

Observe that (7) can be expressed in matrix notation as

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{13}$$

That is, the action of $\mathbf{F}$ on $\mathbf{x}$ is equivalent to multiplication by $\mathbf{A}$, where $\mathbf{A}$ is the 2×3 matrix in (13),

$$\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}. \tag{14}$$

Note that we do not say that "the transformation $\mathbf{F}$ is the matrix $\mathbf{A}$." Rather, we say that $\mathbf{F}$ is the transformation *multiplication by A*. For that reason we call $\mathbf{F}$ a **matrix transformation**. The point that we wish to make here is that the linear transformation given in Example 3 happens to be a matrix transformation. We now show that that correspondence is no accident.

THEOREM 10.8.1 *Matrix Transformation*
A transformation $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if it is a matrix transformation.

Proof: First, we show that if $\mathbf{F}$ is a matrix transformation [i.e., if there is an $m \times n$ matrix $\mathbf{A}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for each $\mathbf{x}$ in $\mathbb{R}^n$], then $\mathbf{F}$ is linear. That is easy since $\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) = \mathbf{A}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\mathbf{A}\mathbf{u} + \beta\mathbf{A}\mathbf{v} = \alpha\mathbf{F}(\mathbf{u}) + \beta\mathbf{F}(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$ and for all scalars α, β. To prove the converse, let $\{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ and $\{\mathbf{j}_1, \ldots, \mathbf{j}_m\}$ be bases for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. We may express

$$\mathbf{x} = \sum_{j=1}^{n} x_j \mathbf{i}_j \quad \text{and} \quad \mathbf{y} = \sum_{k=1}^{m} y_k \mathbf{j}_k.$$

Then

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}\left(\sum_{j=1}^{n} x_j \mathbf{i}_j\right) = \sum_{j=1}^{n} x_j \mathbf{F}(\mathbf{i}_j)$$

by (12). Since $\mathbf{F}(\mathbf{i}_j)$ is in $\mathbb{R}^m$, it can be expressed in the form

$$\mathbf{F}(\mathbf{i}_j) = \sum_{k=1}^{m} a_{kj} \mathbf{j}_k$$

so

$$\mathbf{F}(\mathbf{x}) = \sum_{j=1}^{n} x_j \sum_{k=1}^{m} a_{kj} \mathbf{j}_k = \sum_{k=1}^{m} \left(\sum_{j=1}^{n} a_{kj} x_j\right) \mathbf{j}_k. \tag{15}$$

But we also have

$$\mathbf{F}(\mathbf{x}) = \mathbf{y} = \sum_{k=1}^{m} y_k \mathbf{j}_k, \tag{16}$$

and it follows from (15) and (16), and the linear independence of the $\mathbf{j}_k$'s, that

$$y_k = \sum_{j=1}^{n} a_{kj} x_j$$

or $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} = \{a_{kj}\}$ is $m \times n$, as was to be proved. ■

We now introduce some additional terminology. First, recall that $\mathbf{F}$ is understood to be single valued. That is, to each vector $\mathbf{x}$ in the domain V of $\mathbf{F}$ there corresponds a unique image $\mathbf{F}(\mathbf{x})$ in the range R of $\mathbf{F}$. If, in addition to each vector in R there corresponds a unique inverse image in V, then $\mathbf{F}$ is said to be **one-to-one**. Notice that we do not say "to each vector in W there corresponds a unique inverse image in V" since R may not be all of W, in which case those vectors which are in W but not in R have no inverse image at all. If R does turn out to be all of W, then $\mathbf{F}$ is said to be **onto**; that is, $\mathbf{F}$ maps V "onto" W rather than "into" W. Finally, if $\mathbf{F}$ is both one-to-one and onto, it is said to be **invertible** for then every vector in W has a unique image in V. This inverse transformation, from W onto V, is called the **inverse of F** and is denoted as $\mathbf{F}^{-1}$.

EXAMPLE 5. Consider the matrix transformation $\mathbf{F}$ in Example 2. There, $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, with

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 2 & 2 & -1 \end{bmatrix}. \tag{17}$$

From (5c) we see that R is only the two-dimensional subspace of W ($= \mathbb{R}^3$) consisting of the plane $c_3 - 2c_2 + c_1 = 0$ so that $\mathbf{F}$ is not *onto*. That result is illustrated schematically in Fig. 3, where R is shown to be only a part of W. Furthermore, if $\mathbf{c}$ is in R [i.e., if (5c) is satisfied], then (5) yields a nonunique solution for $\mathbf{x}$, so $\mathbf{F}$ is not one-to-one. Summarizing, $\mathbf{F}$ is neither one-to-one nor onto and is therefore not invertible. ■

EXAMPLE 6. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 2 & -1 \end{bmatrix}. \tag{18}$$

Applying elementary operations to the system $\mathbf{A}\mathbf{x} = \mathbf{c}$, namely, to

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, \tag{19}$$

we obtain the equivalent system

$$\begin{aligned} x_1 - 2x_2 &= c_1, \\ 3x_2 &= c_2 - c_1, \\ 0 &= c_3 - c_2 - c_1, \end{aligned} \tag{20a,b,c}$$

from which it is seen that $\mathbf{F}$ is *not* onto (why not?), although it is one-to-one (why?). Thus $\mathbf{F}$ is not invertible. ▮

EXAMPLE 7. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & -3 \end{bmatrix}. \tag{21}$$

Applying elementary operations to the system $\mathbf{Ax} = \mathbf{c}$, namely, to

$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{22}$$

we obtain the equivalent system

$$\begin{aligned} x_1 + \quad x_2 + 2x_3 &= c_1, \\ 2x_2 + 7x_3 &= 2c_1 - c_2, \end{aligned} \tag{23a,b}$$

from which it is seen that $\mathbf{F}$ *is* onto (why?), although *not* one-to-one (why not?). Thus, $\mathbf{F}$ is not invertible. ▮

EXAMPLE 8. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{24}$$

Applying elementary operations to the system $\mathbf{Ax} = \mathbf{c}$, we obtain

$$\begin{aligned} 2x_1 - \quad x_2 + \quad x_3 &= c_1, \\ 3x_2 + \quad x_3 &= c_2, \\ - 2x_3 &= 2c_3 - c_2 - c_1, \end{aligned} \tag{25a,b,c}$$

from which it is seen that $\mathbf{F}$ is both onto (so that $\mathbf{F}$ is an operator) and one-to-one, and is therefore invertible; completion of the Gauss–Jordan reduction of (25) reveals that the inverse operator $\mathbf{F}^{-1}$ is the matrix operator associated with the matrix

$$\begin{bmatrix} \frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ -\frac{1}{6} & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix}, \tag{26}$$

which, of course, is the inverse of the $\mathbf{A}$ matrix, $\mathbf{A}^{-1}$. ▮

Closure. Recall that we develop n-space in Chapter 9, and then generalize the vector space concept in Section 9.6. The role of the present section is analogous in that here we have generalize the concept of matrix, developed earlier in this chapter, to that of *transformations* on vector spaces.

Besides establishing the concept, together with the standard mathematical terminology, the key result is given in Theorem 10.8.1, that a transformation $\mathbf{F}: \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if it is a matrix transformation.

EXERCISES 10.8

1. In general, the effect of a transformation on the input vector varies from one input vector to another. For example, let $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2$ be a matrix transformation $\mathbf{F}(\mathbf{x}) = \mathbf{Ax}$, where $\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}$. In 2-space the "effect" of a transformation on a nonzero vector $\mathbf{x}$ amounts to the resulting rotation in the plane, and the dilation (i.e., $\|\mathbf{Ax}\| \,/\, \|\mathbf{x}\|$). For the transformation $\mathbf{F}$ given above, show that these effects are as follows for the given input vectors, and notice that the effect of $\mathbf{F}$ varies from one $\mathbf{x}$ to another.

	Input	Effect of $\mathbf{F}$
(a)	$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$	rotation $= 0$ radians dilation $= 3$
(b)	$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	rotation ≈ 1.1 radians dilation $= \sqrt{5}$
(c)	$\mathbf{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$	rotation $= 0$ radians dilation $= 1$

2. Determine whether $\mathbf{F}$ is linear or nonlinear by determining whether or not the deviation $\mathbf{F}(\alpha\,\mathbf{u} + \beta\,\mathbf{v}) - \alpha\mathbf{F}(\mathbf{u}) - \beta\mathbf{F}(\mathbf{u})$ is necessarily zero.

(a) $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_1 + x_2 \end{bmatrix}$

(b) $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} 3x_1 \\ x_1 + x_2 \end{bmatrix}$

(c) $\mathbf{F}: \mathbb{R}^3 \to \mathbb{R}^2, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1x_2 \\ x_3 \end{bmatrix}$

(d) $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^1, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} 3x_2 \end{bmatrix}$

(e) $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} \sin x_2 \\ 4x_1 \end{bmatrix}$

(f) $\mathbf{F}: \mathbb{R}^2 \to \mathbb{R}^2, \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 + 1 \\ x_2 + 1 \end{bmatrix}$

3. (*Identity operator*) We say that $\mathbf{I}: V \to V$ is an **identity operator** if $\mathbf{I}(\mathbf{x}) = \mathbf{x}$ for each $\mathbf{x}$ in V.

(a) Suppose that $\mathbf{F}: V \to V$ is linear and that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis for V. Show that if $\mathbf{F}(\mathbf{v}_1) = \mathbf{v}_1, \dots, \mathbf{F}(\mathbf{v}_n) = \mathbf{v}_n$, then $\mathbf{F} = \mathbf{I}$.

(b) Determine the matrix $\mathbf{A}$ corresponding to the identity operator $\mathbf{I}: \mathbb{R}^n \to \mathbb{R}^n$, i.e., such that $\mathbf{I}(\mathbf{x}) = \mathbf{Ax}$.

4. (*Zero transformation*) We say that $\mathbf{\Phi}: V \to W$ is a **zero transformation** if $\mathbf{\Phi}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$'s in V.

(a) Suppose that $\mathbf{F}: V \to W$ is linear and that $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis for V. Show that if $\mathbf{F}(\mathbf{v}_1) = \mathbf{0}, \dots, \mathbf{F}(\mathbf{v}_n) = \mathbf{0}$, then $\mathbf{F} = \mathbf{\Phi}$.

(b) Determine the matrix $\mathbf{A}$ corresponding to the zero transformation $\mathbf{\Phi}: \mathbb{R}^n \to \mathbb{R}^m$, i.e., such that $\mathbf{\Phi}(\mathbf{x}) = \mathbf{Ax}$.

5. In each case $\mathbf{F}: \mathbb{R}^n \to \mathbb{R}^m$ is the matrix transformation corresponding to the given $m \times n$ matrix $\mathbf{A}$. Determine dim$R$, dim$K$, and dim$V$. Is $\mathbf{F}$ onto? One-to-one? Invertible? Explain. Put forward any basis for K and any basis for R (if, indeed, they have bases; see Exercise 7 in Section 9.9).

(a) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 4 & 3 & 3 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 4 & 1 \\ 3 & 2 \\ 0 & -1 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 6 & 0 \\ 0 & -2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & -1 & 1 \\ 0 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$ (h) $\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 0 & 1 \\ 0 & 3 \end{bmatrix}$

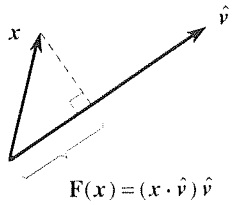
6. Make up any example of a matrix transformation $\mathbf{F}: \mathbb{R}^n \to \mathbb{R}^m$ that is one-to-one but not onto, one that is onto but not one-to-one, one that is neither one-to-one nor onto, and one that is both one-to-one and onto. If such an example is impossible, explain why that is so.

(a) $n = 2, m = 2$ (b) $n = 3, m = 2$
(c) $n = 2, m = 3$ (d) $n = 3, m = 3$

7. (*Projection operators*) Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ be the transformation

$$\mathbf{F}(\mathbf{x}) = (\mathbf{x} \cdot \hat{\mathbf{v}})\hat{\mathbf{v}}, \tag{7.1}$$

where $\hat{\mathbf{v}} = [v_1, v_2, v_3]^{\mathrm{T}}$ is a prescribed unit vector. In geometric terms, $\mathbf{F}(\mathbf{x})$ is the vector orthogonal projection of $\mathbf{x}$ onto the line of action of $\hat{\mathbf{v}}$, as illustrated below. Hence, $\mathbf{F}$ in (7.1) is known as a **projection operator**. We now state the problem: Show that $\mathbf{F}$ is linear, so that one can express $\mathbf{F}(\mathbf{x})$ as $\mathbf{Ax}$. Then determine the nine elements of the $\mathbf{A}$ matrix. (They will depend on v_1, v_2, v_3.)



$\mathbf{F}(x) = (x \cdot \hat{v})\,\hat{v}$

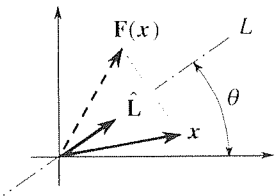
8. (*More about projection operators*) Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ be the transformation

$$\mathbf{F}(\mathbf{x}) = (\mathbf{x} \cdot \hat{\mathbf{v}}_1)\hat{\mathbf{v}}_1 + (\mathbf{x} \cdot \hat{\mathbf{v}}_2)\hat{\mathbf{v}}_2, \tag{8.1}$$

where $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2$ are prescribed ON vectors. Then $\mathbf{F}(\mathbf{x})$ is the vector orthogonal projection of $\mathbf{x}$ onto the *plane* spanned by $\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2\}$.
(a) Given that $\hat{\mathbf{v}}_1 = [1, 0, 0]^{\mathrm{T}}$ and $\hat{\mathbf{v}}_2 = [0, 1, 0]^{\mathrm{T}}$, work out $\mathbf{F}(\mathbf{x})$ for $\mathbf{x} = [2, 3, 4]^{\mathrm{T}}$, and draw an informative, labeled picture, analogous to the one shown in Exercise 7.
(b) Show that $\mathbf{F}$ is linear so that one can express $\mathbf{F}(\mathbf{x})$ as $\mathbf{Ax}$. Then determine the nine elements of the $\mathbf{A}$ matrix, in terms of the components v_{11}, v_{12}, v_{13} of $\hat{\mathbf{v}}_1$ and v_{21}, v_{22}, v_{23} of $\hat{\mathbf{v}}_2$.

9. Show that $\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\mathbf{F}(\mathbf{u}) + \beta\mathbf{F}(\mathbf{v})$, in Definition 10.8.1, is equivalent to the two conditions $\mathbf{F}(\mathbf{u} + \mathbf{v}) = \mathbf{F}(\mathbf{u}) + \mathbf{F}(\mathbf{v})$ and $\mathbf{F}(\alpha\mathbf{u}) = \alpha\mathbf{F}(\mathbf{u})$.

10. (*Reflection about a line*) Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ reflect any given vector $\mathbf{x}$ about the line L, as shown in the accompanying figure.



(a) Show that $\mathbf{F}(\mathbf{x}) = \mathbf{x} + 2[(\mathbf{x} \cdot \hat{\mathbf{L}})\hat{\mathbf{L}} - \mathbf{x}]$.
(b) Show that $\mathbf{F}$ is linear and determine the matrix $\mathbf{A}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{Ax}$.
(c) Work out $\mathbf{A}^2$ and show that $\mathbf{A}^2 = \mathbf{I}$ (Exercise 3). Why should it have been obvious, without working out $\mathbf{A}^2$, that $\mathbf{A}^2 = \mathbf{I}$?

11. (*Linear combination and composition*) If $\mathbf{F}$ and $\mathbf{G}$ are transformations from V into W, then we define the **linear combination** of $\mathbf{F}$ and $\mathbf{G}$, $(\alpha\mathbf{F} + \beta\mathbf{G}) : V \to W$, by

$$(\alpha\mathbf{F} + \beta\mathbf{G})(\mathbf{x}) \equiv \alpha\mathbf{F}(\mathbf{x}) + \beta\mathbf{G}(\mathbf{x}) \tag{11.1}$$

for all $\mathbf{x}$ in V. Given transformations $\mathbf{F} : U \to V$ and $\mathbf{G} : V \to W$, we define the **composition** of $\mathbf{F}$ and $\mathbf{G}$, $(\mathbf{GF}) : U \to W$, by

$$(\mathbf{GF})(\mathbf{x}) \equiv \mathbf{G}(\mathbf{F}(\mathbf{x})) \tag{11.2}$$

for all $\mathbf{x}$ in U.

(a) Let $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^2$ and $\mathbf{G} : \mathbb{R}^2 \to \mathbb{R}^2$ be matrix transformations with matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 3 & -2 \\ 2 & 5 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 4 & 0 \end{bmatrix},$$

respectively. Evaluate $(\mathbf{GF})(\mathbf{x})$ for $\mathbf{x} = [3, 1, -2, 6]^{\mathrm{T}}$. Find a single matrix corresponding to the composite transformation $\mathbf{GF}$.
(b) Let $\mathbf{F} : V \to V$ be a linear operator, and define the composite transformation $\mathbf{F}^2 : V \to V$ by $\mathbf{F}^2(\mathbf{x}) \equiv \mathbf{F}(\mathbf{F}(\mathbf{x}))$. Show that $\mathbf{F}^2$ is linear, too.

12. Show that the *translation operator* $\mathbf{F}(\mathbf{x}) = \mathbf{x} + \mathbf{c}$, where $\mathbf{c}$ is a constant vector, is nonlinear.

13. (*Applications to computer graphics*) It is basic, in computer graphics, to be able to move points about, in 3-space, by combinations of translations and rotations. Translation is easily accomplished by the operator

$$\mathbf{F}(\mathbf{X}) = \mathbf{X} + \Delta\mathbf{X}, \tag{13.1}$$

where $\mathbf{X} = [x, y, z]^{\mathrm{T}}$ is the position vector to the point and $\Delta\mathbf{X} = [\Delta x, \Delta y, \Delta z]^{\mathrm{T}}$ is the translation. However, whereas it is convenient, in the computer software, to express all translations and rotations as matrix transformations, the operator $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ is not linear (Exercise 12), and hence not expressible as a matrix transformation. To circumvent this difficulty we define $\mathbf{X} = [x, y, z, 1]^{\mathrm{T}}$, instead, where the fourth component, unity, is included for convenience. Then we can express

$$\mathbf{F}(\mathbf{X}) = \mathbf{TX}$$
$$= \begin{bmatrix} 1 & 0 & 0 & \Delta x \\ 0 & 1 & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x + \Delta x \\ y + \Delta y \\ z + \Delta z \\ 1 \end{bmatrix}, \tag{13.2}$$

which, if we pay attention to only the first three components, effects the translation by means of multiplication by $\mathbf{T}$, where $\mathbf{T}$ is the 4×4 matrix in (13.2).

(a) Show that

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_z\mathbf{X} = \begin{bmatrix} c_z & -s_z & 0 & 0 \\ s_z & c_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{13.3}$$

effects a rotation about the z axis through an angle θ_z, taken according to the right-hand rule, where c_z, s_z are shorthand for $\cos\theta_z$, $\sin\theta_z$, respectively. HINT: Letting $x = r\cos\theta$, $y = r\sin\theta$, show that

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_z\mathbf{X} = \begin{bmatrix} r\cos(\theta + \theta_z) \\ r\sin(\theta + \theta_z) \\ z \\ 1 \end{bmatrix}.$$

NOTE: Similarly, rotations about the x axis through an angle θ_x, and about the y axis through an angle θ_y, are effected by

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_y\mathbf{X} = \begin{bmatrix} c_y & 0 & -s_y & 0 \\ 0 & 1 & 0 & 0 \\ s_y & 0 & c_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{13.4}$$

and

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_x\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_x & -s_x & 0 \\ 0 & s_x & c_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{13.5}$$

respectively, where c_x, s_x, c_y, s_y denote $\cos\theta_x, \sin\theta_x, \cos\theta_y, \sin\theta_y$, respectively.

(b) Show that a rotation about the z axis, followed by a translation, is effected by the composite transformation (see Exercise 11)

$$\mathbf{F}(\mathbf{X}) = \mathbf{TR}_z\mathbf{X} = \begin{bmatrix} c_z & -s_z & 0 & \Delta x \\ s_z & c_z & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}.$$

Does the order of the operations matter? That is, is $\mathbf{TR}_z = \mathbf{R}_z\mathbf{T}$?

(c) Compute $\mathbf{F}(\mathbf{X}) = \mathbf{TR}_x\mathbf{R}_y\mathbf{R}_z\mathbf{X}$ for $\mathbf{X} = [1, 1, 0, 1]^{\mathrm{T}}$, $\theta_z = -\pi/4$, $\theta_y = \pi/2$, $\theta_x = \pi$, $\Delta x = 2$, $\Delta y = 1$, $\Delta z = -3$. Verify the result by drawing the coordinate axes, identifying the initial point $\mathbf{X}$, and then carrying out each rotation and translation graphically in a neat sketch.

(d) Let the point and eraser of a pencil be located, initially, by $\mathbf{X}_p = [0, 1, 1, 1]^{\mathrm{T}}$ and $\mathbf{X}_e = [0, 1, 0, 1]^{\mathrm{T}}$, respectively. Locate the point and the eraser following the composite transformation

$$\mathbf{F}(\mathbf{X}) = \mathbf{TR}_x\mathbf{R}_y\mathbf{R}_z\mathbf{X},$$

where $\theta_x = 0.2$, $\theta_y = 0.3$, $\theta_z = -0.6$, $\Delta x = 1$, $\Delta y = 3$, $\Delta z = 1$. In a neat sketch, show the pencil in its initial and final configurations, and verify that its length has remained the same.

(e) Repeat part (d), with $\theta_x = \pi/2$, $\theta_y = 0$, $\theta_z = -\pi$, $\Delta x = \Delta y = \Delta z = 1$.

(f) Repeat part (d), with $\theta_x = -\pi/2$, $\theta_y = \theta_z = \pi/2$, $\Delta x = \Delta y = 1$, $\Delta z = 0$.

Chapter 10 Review

The following review is limited to a number of isolated results and formulas that should be both understood and memorized.

Matrix multiplication:

$$\mathbf{AB} \neq \mathbf{BA} \qquad \text{(in general)}$$

Transpose:

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

Determinants:

$$\begin{aligned} \det(\alpha\mathbf{A} + \beta\mathbf{B}) &\neq \alpha \det\mathbf{A} + \beta \det\mathbf{B} \qquad \text{(in general)} \\ \det(\mathbf{AB}) &= (\det\mathbf{A})(\det\mathbf{B}) \end{aligned}$$

Rank:

$$\begin{aligned} r(\mathbf{A}) &= \text{number of LI columns in } \mathbf{A} \\ &= \text{number of LI rows in } \mathbf{A} \end{aligned}$$

Systems of linear algebraic equations, $\mathbf{Ax} = \mathbf{c}$ (where $\mathbf{A}$ is $m \times n$):

Inconsistent:

$$\text{No solution if} \quad r(\mathbf{A}|\mathbf{c}) \neq r(\mathbf{A})$$

Consistent:

$$\begin{aligned} &\text{Unique solution if} \quad r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n \\ &(n - r)\text{-parameter family of solutions if} \quad r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = r < n \end{aligned}$$

The case where $m = n$:

$$\text{Unique solution} \quad \left(\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}\right) \text{ if and only if } \det\mathbf{A} \neq 0 \quad [\text{i.e., } r(\mathbf{A}) = n]$$

Inverse matrix, $\mathbf{A}^{-1}$:

Exists, and is unique, if and only if $\det\mathbf{A} \neq 0$.

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{A} &= \mathbf{AA}^{-1} = \mathbf{I} \\ (\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} \\ (\mathbf{A}^{-1})^{\mathrm{T}} &= (\mathbf{A}^{\mathrm{T}})^{-1} \end{aligned}$$

Orthogonal matrices:

A matrix $\mathbf{Q}$ is orthogonal if it is square and its columns are ON.

Chapter 11

The Eigenvalue Problem

11.1 Introduction

In this chapter we study the problem

$$\boxed{\mathbf{Ax} = \lambda\mathbf{x},} \tag{1}$$

where $\mathbf{A}$ is a given $n \times n$ matrix, $\mathbf{x}$ is an unknown $n \times 1$ vector, and λ is an unknown scalar. If we re-express (1) as $\mathbf{Ax} = \lambda\mathbf{Ix}$ (where $\mathbf{I}$ is an $n \times n$ identity matrix), then subtraction of $\lambda\mathbf{Ix}$ from both sides gives the equivalent equation*

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}, \tag{2}$$

which is a homogeneous system of n equations in the n unknown x_j's, where the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ contains the parameter λ.

To be sure that (1) and (2) are clear, let us write them out in scalar form, for $n = 3$, for example. Then (1) is the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= \lambda x_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= \lambda x_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= \lambda x_3. \end{aligned}$$

Subtracting the terms on the right from those on the left gives

$$\begin{aligned} (a_{11} - \lambda)x_1 + \qquad a_{12}x_2 + \qquad a_{13}x_3 &= 0, \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \qquad a_{23}x_3 &= 0, \\ a_{31}x_1 + \qquad a_{32}x_2 + (a_{33} - \lambda)x_3 &= 0, \end{aligned}$$

which, in matrix form, is equation (2).

*Of course we don't need to insert the $\mathbf{I}$. We could re-express (1), correctly, as $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0}$, but it would *not* follow from the latter that $(\mathbf{A} - \lambda)\mathbf{x} = \mathbf{0}$ because subtraction of a scalar (λ) from a matrix ($\mathbf{A}$) is not defined. Hence the need to insert $\mathbf{I}$.

From Chapter 10, we know that (2) is consistent because it necessarily admits the "trivial" solution $\mathbf{x} = \mathbf{0}$. However, our interest in (2) shall be in the search for *nontrivial* solutions, and we anticipate that whether or not nontrivial solutions exist will depend upon the value of λ. Thus, the problem of interest is as follows: given the $n \times n$ matrix $\mathbf{A}$, find the value(s) of λ (if any) such that (2) admits nontrivial solutions, and find those nontrivial solutions. The latter is called the **eigenvalue problem** and is the focus of this chapter. The λ's that lead to nontrivial solutions for $\mathbf{x}$ are called the **eigenvalues** (or **characteristic values**), and the corresponding nontrivial solutions for $\mathbf{x}$ are called the **eigenvectors** (or **characteristic vectors**).

The eigenvalue problem (1) [or, equivalently, (2)] occurs in a wide variety of applications such as vibration theory, chemical kinetics, stability of equilibria, buckling of structures, convergence of iterative techniques, and systems of coupled ordinary differential equations. To place the eigenvalue problem (1) in perspective, recall that in Chapter 8 and 10 we studied the problem

$$\mathbf{A}\mathbf{x} = \mathbf{c} \tag{3}$$

of m linear algebraic equations in n unknowns (i.e., $\mathbf{A}$ was $m \times n$). In general, $\mathbf{c} \neq \mathbf{0}$, in which case (3) was said to be nonhomogeneous. The eigenvalue problem is by no means unrelated to (3); it amounts to a special case, where $\mathbf{c} = \mathbf{0}$ (i.e., it is homogeneous), where $m = n$, and where the coefficient matrix "$\mathbf{A}$" $= \mathbf{A} - \lambda\mathbf{I}$ contains the parameter λ. Thus, to solve the eigenvalue problem we will be able to use results already established in Chapter 10.

11.2 Solution Procedure and Applications

11.2.1. Solution and applications. The eigenvalue problem

$$\boxed{(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}} \tag{1}$$

has the unique trivial solution $\mathbf{x} = \mathbf{0}$ if $\det(\mathbf{A} - \lambda\mathbf{I}) \neq 0$, and nontrivial solutions (in addition to the trivial solution) if and only if

$$\boxed{\det(\mathbf{A} - \lambda\mathbf{I}) = 0.} \tag{2}$$

The latter is not a vector or matrix equation; it is an algebraic equation in λ, known as the **characteristic equation** corresponding to the matrix $\mathbf{A}$, and its left-hand side is an nth degree polynomial known as the **characteristic polynomial**. According to the fundamental theorem of algebra, such an equation has precisely n roots in the complex plane. Since one or more of these roots can be repeated, we can say that there is at least one eigenvalue λ, and at most n distinct eigenvalues λ, corresponding to any given $n \times n$ matrix $\mathbf{A}$.

As in Chapter 10, we continue to consider only real matrices. However, even if $\mathbf{A}$ is real (so that the coefficients of the characteristic polynomial are too), the characteristic equation can still have complex roots. That case will not be very important to us. Thus, we avoid it entirely in Chapter 11, and consider it briefly in Chapter 12.

This is not the first time we have run into the need to solve polynomial equations. In Section 3.4 we sought solutions to linear, homogeneous, constant-coefficient differential equations by seeking $y(x) = e^{\lambda x}$. Putting that solution form into the nth order differential equation gave an nth degree polynomial equation on λ. In fact, even the terminology was the same: the equation was called the **characteristic equation** of the differential equation, and the nth degree polynomial was called the **characteristic polynomial**. If $n = 2$ we can solve the characteristic equation by the quadratic formula. For larger n's we can, if necessary, use computer software such as the *Maple* fsolve command discussed in Section 3.4. Thus, let us consider (2) to have been solved for the eigenvalues, for the moment, and let us designate them as $\lambda_1, \ldots, \lambda_k$ $(1 \le k \le n)$.

Next, set $\lambda = \lambda_1$ in (1). Since $\det(\mathbf{A} - \lambda_1\mathbf{I}) = 0$, it is guaranteed that $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ will have nontrivial solutions. We can find those solutions by Gauss elimination, and we designate them as $\mathbf{e}_1$, where the letter e is for eigenvector. The $\mathbf{e}_1$ solution space is called the **eigenspace** corresponding to the eigenvalue λ_1.* Next, we set $\lambda = \lambda_2, \ldots, \lambda_k$ and repeat the process until the k eigenspaces have been found.

EXAMPLE 1. Determine all eigenvalues and eigenspaces of

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}. \tag{3}$$

The characteristic equation is

$$\begin{aligned}\det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} 2-\lambda & 2 & 1 \\ 1 & 3-\lambda & 1 \\ 1 & 2 & 2-\lambda \end{vmatrix} = \lambda^3 - 7\lambda^2 + 11\lambda - 5 \\ &= (\lambda - 5)(\lambda - 1)^2 = 0 \end{aligned} \tag{4}$$

so the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 5$ and $\lambda_2 = 1$ (or vice versa since the order is immaterial), with $\lambda_2 = 1$ called a *repeated eigenvalue* – specifically, an eigenvalue of *multiplicity* 2 because it is a double root of the characteristic equation (4).

Next, find the eigenspaces.

$\lambda_1 = 5$: Then $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ becomes

$$\begin{bmatrix} 2-5 & 2 & 1 \\ 1 & 3-5 & 1 \\ 1 & 2 & 2-5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 2 & -3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{5}$$

*Note that the eigenspace corresponding to an eigenvalue λ_j is not quite the same as the set of eigenvectors corresponding to λ_j, it is that set *plus* the trivial solution (which is NOT itself an eigenvector). The reason we define the eigenspace corresponding to λ_j as the entire solution space of $(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{x} = \mathbf{0}$ (i.e., including the zero solution) is so that the eigenspace will be a *vector space*, for recall that a vector space must contain a zero vector.

Gauss elimination of which gives

$$\begin{bmatrix} -3 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

The solution is $x_3 = \alpha$ (arbitrary), $x_2 = \alpha$, $x_1 = \alpha$ so, using $\mathbf{e}$ in place of $\mathbf{x}$,

$$\mathbf{e} = \begin{bmatrix} \alpha \\ \alpha \\ \alpha \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{7}$$

Thus, the eigenspace corresponding to $\lambda_1 = 5$ is span$\{[1, 1, 1]^{\mathrm{T}}\}$, the latter being a one-dimensional subspace of $\mathbb{R}^3$, namely, the line through the origin given by (7).

$\lambda_2 = 1$: Then $(\mathbf{A} - \lambda_2 \mathbf{I})\mathbf{x} = \mathbf{0}$ becomes

$$\begin{bmatrix} 2-1 & 2 & 1 \\ 1 & 3-1 & 1 \\ 1 & 2 & 2-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{8}$$

Gauss elimination of which gives

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{9}$$

The solution is $x_3 = \beta$ (arbitrary), $x_2 = \gamma$ (arbitrary), $x_1 = -\beta - 2\gamma$ so

$$\mathbf{e} = \begin{bmatrix} -\beta - 2\gamma \\ \gamma \\ \beta \end{bmatrix} = \beta \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}. \tag{10}$$

Thus, the eigenspace corresponding to $\lambda_2 = 1$ is span$\{[-1, 0, 1]^{\mathrm{T}}, [-2, 1, 0]^{\mathrm{T}}\}$, the latter being a two-dimensional subspace of $\mathbb{R}^3$, namely, the plane through the origin, spanned by $[-1, 0, 1]^{\mathrm{T}}$ and $[-2, 1, 0]^{\mathrm{T}}$. In fact, the equation of that plane is seen, in (9), as $x_1 + 2x_2 + x_3 = 0$.

COMMENT 1. We can determine eigenvectors only up to arbitrary scale factors [such as the α in (7)] because, if a vector $\mathbf{e}$ satisfies $\mathbf{Ae} = \lambda\mathbf{e}$, then so does any scalar multiple of $\mathbf{e}$. Along these lines, observe that it would be correct to write $\mathbf{e} = \beta[1, 0, -1]^{\mathrm{T}} + \gamma[-2, 1, 0]^{\mathrm{T}}$, say, since the scale factor of -1 in the first vector can be absorbed by the arbitrary β.

COMMENT 2. In the language of Section 10.5, the rank of the $\mathbf{A} - \lambda_1\mathbf{I}$ coefficient matrix in (6) is 2 so $n - r = 3 - 2 = 1$, and (6) admits a one-parameter family of solutions. That is, the nullity of $\mathbf{A} - \lambda_1\mathbf{I}$ is 1 and the $\mathbf{e}_1$ eigenspace is one-dimensional. Similarly, the rank of $\mathbf{A} - \lambda_2\mathbf{I}$ in (9) is 1, so $n - r = 3 - 1 = 2$, and (9) admits a two-parameter family of solutions. That is, the nullity of $\mathbf{A} - \lambda_2\mathbf{I}$ is 2, and the $\mathbf{e}_2$ eigenspace is two-dimensional. However, there is no reason to believe that the multiplicity of an eigenvalue necessarily equals the dimension of the corresponding eigenspace, even though it happens to be true in this example. ▮

EXAMPLE 2. The matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

has the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 0 & 1 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^3 = 0 \tag{12}$$

with roots $\lambda = 1, 1, 1$. That is, $\lambda_1 = 1$ is a root of multiplicity three. To find the eigenspace, write out $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ as

$$\begin{bmatrix} 1-1 & 0 & 1 \\ 1 & 1-1 & 0 \\ 0 & 0 & 1-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{13}$$

The solution is $x_3 = 0$, $x_1 = 0$, $x_2 = \alpha$ (arbitrary) so

$$\mathbf{e} = \begin{bmatrix} 0 \\ \alpha \\ 0 \end{bmatrix} = \alpha \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{14}$$

Thus, in this case an eigenvalue of multiplicity three gave rise to an eigenspace of dimension one. ▮

With the "mechanics" of the eigenvalue problem explained in the first two examples, let us devote the next examples to applications.

EXAMPLE 3. *Solution of Differential Equations.* How can we solve the coupled differential equations

$$\begin{aligned} x' &= x + 4y, \\ y' &= x + y \end{aligned} \tag{15}$$

on $x(t)$ and $y(t)$? We could use the method of elimination (Section 3.9) to uncouple them, or solve by the Laplace transform method (Chapter 5). Here, we pursue a different approach, that will lead to an eigenvalue problem.

Since (15) is linear, constant-coefficient, and homogeneous, we can find exponential solutions. Thus, seek x, y in the form

$$x(t) = q_1 e^{rt}, \quad y(t) = q_2 e^{rt}, \tag{16}$$

where q_1, q_2, r are constants that are to be determined. Putting (16) into (15) gives

$$\begin{aligned} rq_1e^{rt} &= q_1e^{rt} + 4q_2e^{rt}, \\ rq_2e^{rt} &= q_1e^{rt} + q_2e^{rt}, \end{aligned} \tag{17}$$

or, cancelling the e^{rt} 's (because they are nonzero) and expressing the result in matrix form, gives

$$\begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = r \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$$

or

$$\mathbf{A}\mathbf{q} = r\mathbf{q}, \tag{18}$$

which is an eigenvalue problem with $\lambda = r$.

We are not interested in the trivial solution $\mathbf{q} = \mathbf{0}$ because it gives the trivial particular solution $x(t) = y(t) = 0$, of (15), whereas we seek the general solution.

Proceeding as above, we obtain these eigenvalues and eigenspaces:

$$\lambda_1 = 3,\ \mathbf{e}_1 = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -1,\ \mathbf{e}_2 = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{19}$$

Denoting $\mathbf{x}(t) = [x(t), y(t)]^{\mathrm{T}}$, each "eigenpair" gives a solution of (15) as

$$\mathbf{x}(t) = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{3t} \quad \text{and} \quad \mathbf{x}(t) = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-t}. \tag{20}$$

By the linearity of (15), we can superimpose these solutions and thereby obtain the general solution

$$\mathbf{x}(t) = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{3t} + \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-t} \tag{21}$$

or, in scalar form,

$$\begin{aligned} x(t) &= 2\alpha e^{3t} - 2\beta e^{-t}, \\ y(t) &= \alpha e^{3t} + \beta e^{-t}. \end{aligned} \tag{22}$$

Of course, α and β are the integration constants (usually denoted as A, B or C_1, C_2 in the ODE chapters).

COMMENT 1. Since we use q_1 and q_2 in (16), it would be natural to wonder why we don't also allow for different exponents, and seek $x(t) = q_1 \exp(r_1 t)$ and $y(t) = q_2 \exp(r_2 t)$. The reason is that unless $r_1 = r_2$ we obtain only the trivial solution $q_1 = q_2 = 0$ (Exercise 1).

COMMENT 2. The method illustrated in this example can be used for any system of coupled, linear, constant-coefficient homogeneous differential equations. However, it will fail to produce a general solution if $\mathbf{A}$ has a repeated eigenvalue of multiplicity k if the dimension of the corresponding eigenspace is less than k. ▮

EXAMPLE 4. *Markov Process.* Suppose that there is a population exchange between Delaware, Maryland, and Pennsylvania such that, each year, 20% of Delaware's residents move to Maryland and 8% move to Pennsylvania; 12% of Maryland's residents move to

Delaware and 10% to Pennsylvania; 10% of Pennsylvania's residents move to Delaware and 3% move to Maryland. For simplicity, let us ignore gains in population due to births and losses due to deaths – or, equivalently, suppose that these effects are nonzero, but equal and opposite, so as to cancel. Further, let us suppose that the three states are a closed system; that is, they exchange populations only among themselves.

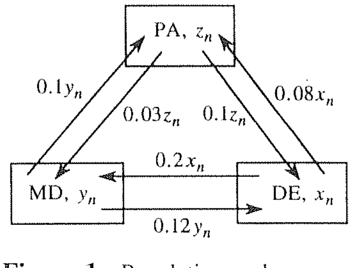


Figure 1. Population exchange model.

If we denote the populations at the end of the nth year, in DE, MD, and PA as x_n, y_n, z_n, respectively, then (see Fig. 1)

$$\begin{aligned} x_{n+1} &= x_n - (0.2 + 0.08)x_n + 0.12y_n + 0.1z_n, \\ y_{n+1} &= y_n + 0.2x_n - (0.12 + 0.1)y_n + 0.03z_n, \\ z_{n+1} &= z_n + 0.08x_n + 0.1y_n - (0.1 + 0.03)z_n, \end{aligned} \tag{23}$$

which are coupled *difference equations*. Difference equations were studied in Section 6.5.3 within the context of differential equations. Here, however, let us consider (23) as a matrix equation,

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 0.72 & 0.12 & 0.1 \\ 0.2 & 0.78 & 0.03 \\ 0.08 & 0.1 & 0.87 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \tag{24}$$

or,

$$\mathbf{p}_{n+1} = \mathbf{A}\mathbf{p}_n, \tag{25}$$

where $\mathbf{p}_n = [x_n, y_n, z_n]^{\mathrm{T}}$ is the "population vector."

The first problem that we pose is to find the population $\mathbf{p}_n$ as a function of n, n being essentially a discrete time variable, given some initial population $\mathbf{p}_0$. That's easy, because (25) gives $\mathbf{p}_1 = \mathbf{A}\mathbf{p}_0$, $\mathbf{p}_2 = \mathbf{A}\mathbf{p}_1 = \mathbf{A}(\mathbf{A}\mathbf{p}_0) = \mathbf{A}^2\mathbf{p}_0$, $\mathbf{p}_3 = \mathbf{A}\mathbf{p}_2 = \mathbf{A}(\mathbf{A}^2\mathbf{p}_0) = \mathbf{A}^3\mathbf{p}_0$, and so on, so

$$\mathbf{p}_n = \mathbf{A}^n\mathbf{p}_0. \tag{26}$$

We wonder whether $\mathbf{A}^n\mathbf{p}_0$ keeps changing as n increases, or whether it settles down and approaches an equilibrium (or steady-state) vector, say $\mathbf{P}$. If there is such an equilibrium vector then, by definition of equilibrium, $\mathbf{p}_{n+1} = \mathbf{p}_n = \mathbf{P}$, so (25) becomes $\mathbf{P} = \mathbf{A}\mathbf{P}$. Surely, $\mathbf{P} = \mathbf{0}$ satisfies

$$\mathbf{A}\mathbf{P} = \mathbf{P}, \tag{27}$$

but the interesting question is whether or not there exist *non*trivial $\mathbf{P}$'s. In fact, (27) is an eigenvalue problem with $\lambda = 1$, so we can say that nontrivial equilibrium vectors exist if and only if 1 is an eigenvalue of $\mathbf{A}$. As explained at the end of this section, we can use *Maple* to obtain the following eigenvalues and eigenspaces of $\mathbf{A}$:

$$\lambda_1 = 1,\ \mathbf{e}_1 = \alpha \begin{bmatrix} -0.56 \\ -0.62 \\ -0.82 \end{bmatrix}; \qquad \lambda_2 = 0.77,\ \mathbf{e}_2 = \beta \begin{bmatrix} -0.09 \\ -0.70 \\ 0.79 \end{bmatrix}; \qquad \lambda_3 = 0.60,\ \mathbf{e}_3 = \gamma \begin{bmatrix} -0.62 \\ 0.70 \\ -0.07 \end{bmatrix}. \tag{28}$$

(Actually, *Maple* gave the λ_j's and $\mathbf{e}_j$'s to nine and ten significant figures, respectively, but we have rounded off for brevity.)

Sure enough, $\lambda = 1$ is among the eigenvalues of $\mathbf{A}$ so there is an equilibrium population vector $\mathbf{P}$ given by the corresponding eigenvector

$$\mathbf{P} = \alpha \begin{bmatrix} 0.56 \\ 0.62 \\ 0.82 \end{bmatrix}, \tag{29}$$

where we have absorbed a factor of -1 into the scale factor α, to avoid the appearance of negative populations. If desired, we can compute α by conserving the total population: $0.56\alpha + 0.62\alpha + 0.82\alpha = x_0 + y_0 + z_0$, so $\alpha = (x_0 + y_0 + z_0)/2.0$.*

Finally, it is important to determine whether or not the equilibrium is stable for it will be observed only if it is stable, just as marbles are found in valleys but not on hilltops. To address the question of stability, let us use the set of LI eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, with any nonzero values of α, β, and γ, as a basis for $\mathbb{R}^3$, and expand the initial vector $\mathbf{p}_0$ in terms of that basis as

$$\mathbf{p}_0 = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3. \tag{30}$$

Then (25) gives

$$\begin{aligned}
\mathbf{p}_1 = \mathbf{A}\mathbf{p}_0 &= c_1\mathbf{A}\mathbf{e}_1 + c_2\mathbf{A}\mathbf{e}_2 + c_3\mathbf{A}\mathbf{e}_3 \\
&= c_1\lambda_1\mathbf{e}_1 + c_2\lambda_2\mathbf{e}_2 + c_3\lambda_3\mathbf{e}_3, \\
\mathbf{p}_2 = \mathbf{A}\mathbf{p}_1 &= c_1\lambda_1\mathbf{A}\mathbf{e}_1 + c_2\lambda_2\mathbf{A}\mathbf{e}_2 + c_3\lambda_3\mathbf{A}\mathbf{e}_3 \\
&= c_1\lambda_1^2\mathbf{e}_1 + c_2\lambda_2^2\mathbf{e}_2 + c_3\lambda_3^2\mathbf{e}_3, \\
&\vdots \\
\mathbf{p}_n = \mathbf{A}\mathbf{p}_{n-1} &= c_1\lambda_1^n\mathbf{e}_1 + c_2\lambda_2^n\mathbf{e}_2 + c_3\lambda_3^n\mathbf{e}_3 \\
&= c_1\mathbf{e}_1 + c_2(0.77)^n\mathbf{e}_2 + c_3(0.60)^n\mathbf{e}_3 \\
&\sim c_1\mathbf{e}_1
\end{aligned} \tag{31}$$

as $n \to \infty$, provided that $c_1 \neq 0$. In fact, c_1 cannot be zero because if it were zero then (31) would give $\mathbf{p}_n \to \mathbf{0}$ as $n \to \infty$ and, since the total population is conserved, that could happen only in the uninteresting case where $\mathbf{p}_0 = \mathbf{0}$. Thus, we see from (31) that $\mathbf{p}_n$ inevitably tends to a multiple of $\mathbf{e}_1$, namely, to the equilibrium vector $\mathbf{P}$.

The upshot is that the population history is given by (26), and that $\mathbf{p}_n$ inevitably tends to a unique steady state which is some scalar multiple of $\mathbf{e}_1$, the multiple being fixed by the conservation of the total population.

COMMENT. This example incorporates a number of linear algebra concepts: matrix multiplication in expressing (23) compactly as (25) and in deriving the solution (26) for $\mathbf{p}_n$; the eigenvalue problem in regard to the possibility of a steady-state solution $\mathbf{P}$; and bases and expansions in assessing the stability of that steady state, in (30)–(31). Consider how effective are these linear algebra concepts and methods in providing a systematic approach to solving this problem, especially in determining the stability of the steady state. The same

*Recall that we built into (24) the assumption that any births and deaths cancel, in number, as revealed by adding the three scalar equations in (24), for that step gives $x_{n+1} + y_{n+1} + z_{n+1} = x_n + y_n + z_n$.

approach applies whether the system includes only three states, as in the present example, or 30 states. ▮

The matrix $\mathbf{A}$ in Example 4 is an example of a "Markov" matrix. An $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is called a **Markov** (or **stochastic**) **matrix** if $a_{ij} \geq 0$ for each i, j, and if the elements of each column sum to unity, or if the elements of each row sum to unity. [In the case of the $\mathbf{A}$ given in (24) its columns sum to unity.] It was no coincidence that $\lambda = 1$ was an eigenvalue, in Example 4 since $\lambda = 1$ is an eigenvalue of *every* Markov matrix.

To prove that claim, suppose that $\mathbf{A}$ is a Markov matrix. The value $\lambda = 1$ will be among the eigenvalues of $\mathbf{A}$ if and only if $\mathbf{Ax} = \mathbf{x}$ has nontrivial solutions for $\mathbf{x}$ or, equivalently, if the rows or columns of $\mathbf{A} - \mathbf{I}$ are linearly dependent. Since $\mathbf{A}$ is a Markov matrix, either the elements of each of its columns sum to unity or the elements of each of its rows sum to unity. It follows that either the elements of each of the columns of $\mathbf{A} - \mathbf{I}$ sum to zero (in which case the rows of $\mathbf{A} - \mathbf{I}$ are linearly dependent) or the elements of each of the rows of $\mathbf{A} - \mathbf{I}$ sum to zero (in which case the columns of $\mathbf{A} - \mathbf{I}$ are linearly dependent), or both. Thus, $\mathbf{A} - \mathbf{I}$ is singular and our claim is proved.

11.2.2. Application to elementary singularities in the phase plane. If you studied Chapter 7, you will recall the fundamental role of the *elementary singularities* in the x, y phase plane, where $x(t)$ and $y(t)$ satisfy the linear ODE's

$$\begin{aligned} x' &= ax + by \\ y' &= cx + dy. \end{aligned} \tag{32}$$

For instance, the system (15) is of that form, so let us reconsider the result, given by (21) and (22), in terms of the x, y phase plane. If $\beta = 0$, then $x(t) = 2\alpha e^{3t}$ and $y(t) = \alpha e^{3t}$, so the phase trajectory is the line $y = x/2$, and if $\alpha = 0$ then $x(t) = -2\beta e^{-t}$ and $y(t) = \beta e^{-t}$, so the phase trajectory is the line $y = -x/2$. These are shown in Fig. 2. The directions of the arrows follow from the fact that e^{3t} increases with t and e^{-t} decreases (since the λ's are of opposite sign), and they imply that (15) has a *saddle* at the origin.

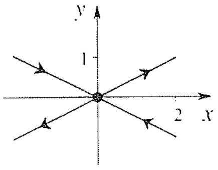


Figure 2. Significance of the eigenvectors.

More generally, observe that we can classify the singularity directly from the eigenvalues of the $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ matrix in (32). Let a, b, c, d be real. Then the eigenvalues are either real or they are complex conjugates. In general, then, we can write $\lambda = \alpha \pm i\beta$, and these λ's contribute solutions of the form $e^{(\alpha \pm i\beta)t}$. We have these possibilities:

λ's real and of the same sign:

λ's < 0 ⇒ stable node

λ's > 0 ⇒ unstable node

λ's real and of opposite sign ⇒ saddle

$$\begin{aligned}&\lambda\text{'s complex}(\lambda = \alpha \pm i\beta):\\ &\qquad \alpha < 0 \;\Rightarrow\; \text{stable focus}\\ &\qquad \alpha = 0 \;\Rightarrow\; \text{center}\\ &\qquad \alpha > 0 \;\Rightarrow\; \text{unstable focus}\end{aligned}$$

In the case of an improper node or saddle, the eigenvectors give the stable and/or unstable manifolds, as mentioned above for the system (15), for which case the stable and unstable manifolds are shown in Fig. 2.

Closure. It is important to remember that the eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$ is *homogeneous* [since it is equivalent to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$] and that the whole point is to find *non*trivial solutions. If, for a given eigenvalue λ, you solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$ by Gauss elimination and obtain $\mathbf{e} = \mathbf{0}$, then your calculations are incorrect: either your eigenvalue is incorrect and/or the Gauss elimination is incorrect.

Observe that our solution strategy uncouples the calculation of the eigenvalues and the eigenvectors: first we solve the characteristic polynomial equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ for the λ's, and then for each λ we solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, by Gauss elimination, for the corresponding eigenvectors.

Computer software. In *Maple*, the relevant commands are **eigenvals** and **eigenvects**, both of which are in the linalg package. The command eigenvals gives just the eigenvalues, and eigenvects gives the eigenvalues, their multiplicity, and a basis for each eigenspace. For instance, let $\mathbf{A}$ be the matrix in Example 1. First, enter

```
with(linalg):
```

and return. Then type

```
A := matrix (3, 3, [2, 2, 1, 1, 3, 1, 1, 2, 2]):
```

because $\mathbf{A}$ is 3×3 and its rows are 2, 2, 1, and 1, 3, 1, and 1, 2, 2, in turn. Then

```
eigenvals(A);
```

gives the eigenvalues as

```
5, 1, 1
```

and

```
eigenvects(A);
```

gives both the eigenvalues and the eigenvectors as

```
[5, 1, {[1, 1, 1]}],    [1, 2, {[−2, 1, 0], [−1, 0, 1]}]
```

In place of the eigenvals command, one can use fsolve to obtain the roots of the characteristic equation, but that is less convenient since to obtain the characteristic equation one needs to expand the $n \times n$ determinant of $\mathbf{A} - \lambda\mathbf{I}$.

Suppose we want the eigenvalues of $\mathbf{A}^{20}$. Enter

```
A := matrix (3, 3, [2, 2, 1, 1, 3, 1, 1, 2, 2]):
```

then

```
evalm(A^20);
```

and then

```
eigenvals(");
```

The quotation mark saves us the trouble of entering the matrix $\mathbf{A}^{20}$ that was calculated in the preceding step.

EXERCISES 11.2

1. (*Example 3*) (a) Derive the eigenvalues and eigenspaces given in (19).
(b) Show that if we assume the forms $x(t) = q_1 \exp(r_1 t)$ and $y(t) = q_2 \exp(r_2 t)$, then we obtain only the trivial solution unless $r_1 = r_2$, as claimed in COMMENT 1.

2. (*Example 4*) To determine the stability of the equilibrium solution $\mathbf{P}$, we expanded $\mathbf{p}_0$ in terms of the basis consisting of the eigenvectors of $\mathbf{A}$. Explain why that choice is particularly convenient. HINT: If necessary, you could try using a different basis, such as $\{[1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}, [0,0,1]^{\mathrm{T}}\}$, instead.

3. Find the eigenvalues and eigenspaces, as well as a basis for each eigenspace.

(a) $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -3 \\ 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} -3 & 2 \\ 6 & -4 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ (h) $\begin{bmatrix} 2 & 1 & 6 \\ 0 & -5 & 3 \\ 0 & 0 & 4 \end{bmatrix}$

(i) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (j) $\begin{bmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 4 \end{bmatrix}$

(k) $\begin{bmatrix} 1 & 0 & 2 \\ 1 & 0 & 2 \\ 1 & 0 & 2 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(m) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}$ (n) $\begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 1 \\ 1 & -13 & 7 \end{bmatrix}$

(o) $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (p) $\begin{bmatrix} 0 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 4 & 0 \end{bmatrix}$

(q) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$ (r) $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix}$

4. (a)–(r) Use computer software to find the eigenvalues and eigenspaces for the matrix in the corresponding part of Exercise 3.

5. Is the following an eigenvector of the matrix $\mathbf{A}$? Explain.

$$\mathbf{A} = \begin{bmatrix} 1 & 8 & 5 & 3 \\ 2 & 16 & 10 & 6 \\ 5 & -14 & -11 & -3 \\ -1 & -8 & -5 & -3 \end{bmatrix}.$$

(a) $[1, 2, -1, 3]^{\mathrm{T}}$ (b) $[1, 2, -4, -1]^{\mathrm{T}}$ (c) $[1, 2, 1, 1]^{\mathrm{T}}$
(d) $[1, 0, 1, -2]^{\mathrm{T}}$ (e) $[1, 0, 1, -1]^{\mathrm{T}}$ (f) $[1, 1, 0, -3]^{\mathrm{T}}$
(g) $[1, 2, -1, -1]^{\mathrm{T}}$ (h) $[2, 1, 0, 1]^{\mathrm{T}}$ (i) $[2, 1, 1, -5]^{\mathrm{T}}$

6. The given matrix has $\lambda = 2$ among its eigenvalues. Find the eigenspace corresponding to that eigenvalue.

(a) $\begin{bmatrix} 3 & 2 & 2 & 1 \\ 2 & 3 & 1 & 2 \\ -1 & 1 & 2 & 0 \\ 2 & 4 & 3 & 5 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 1 & 2 & 1 \\ -1 & 3 & 1 & 2 \\ 0 & 2 & 5 & 3 \\ 1 & 3 & 5 & 6 \end{bmatrix}$

(c) $\begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}$ (d) $\begin{bmatrix} 3 & 0 & 1 & 1 \\ 1 & 3 & 0 & 1 \\ -1 & 1 & 3 & -1 \\ 1 & 2 & 2 & 3 \end{bmatrix}$

7. It is known that the $n \times n$ tridiagonal matrix

$$\mathbf{A} = \begin{bmatrix} b & c & 0 & 0 & & \cdots & 0 \\ a & b & c & 0 & & & \vdots \\ 0 & a & b & c & & & \\ & & & & \ddots & & \vdots \\ \vdots & & & & a & b & c \\ 0 & \cdots & & \cdots & 0 & a & b \end{bmatrix}$$

has eigenvalues

$$\lambda_j = b + 2\sqrt{ac}\,\cos\frac{j\pi}{n+1} \tag{7.1}$$

for $j = 1, 2, \ldots, n$. ($\mathbf{A}$ is called **tridiagonal** because all elements are zero except for those on the main diagonal and the two adjacent diagonals.)

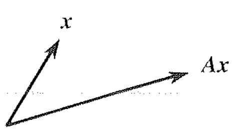
(a) Verify (7.1) by calculating the eigenvalues for $n = 1$ and $n = 2$.
(b) Verify (7.1) by using computer software to determine the eigenvalues for $n = 1$ and 2 and 3.
(c) Verify (7.1) by using computer software to determine the eigenvalues for $n = 4$ and $a = 1, b = 2, c = 1$.
(d) Same as (c), for $n = 4$ and $a = 2, b = 3, c = -1$.
(e) Same as (c), for $n = 5$ and $a = 1, b = 5, c = 3$.

8. Is it possible for a matrix to have no eigenvalues? Explain.

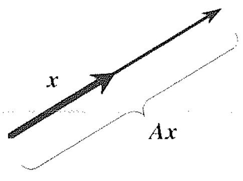
9. We saw in Example 1 that a given eigenvalue can have more than one LI eigenvector. Can a given eigenvector correspond to more than one eigenvalue? Explain.

10. Let $\mathbf{x}$ and $\mathbf{Ax}$ be as shown. Is $\mathbf{x}$ an eigenvector of $\mathbf{A}$? If so, estimate the corresponding eigenvalue; if not, explain why not.

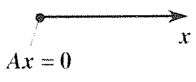
(*a*)



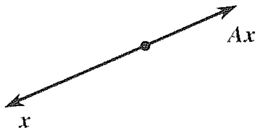
(*b*)



(*c*)



(*d*)



11. Show that the eigenvalues of $k\mathbf{A}$, for any scalar k, are k times those of $\mathbf{A}$. Are the corresponding eigenspaces the same? Explain.

12. Show that the eigenvalues of $\mathbf{A}^{\mathrm{T}}$ are the same as those of $\mathbf{A}$. Is the eigenspace corresponding to an eigenvalue λ of $\mathbf{A}$ the same as the eigenspace corresponding to the same eigenvalue λ of $\mathbf{A}^{\mathrm{T}}$? Prove or disprove.

13. If λ, $\mathbf{e}$ are an eigenvalue and corresponding eigenvector of a matrix $\mathbf{A}$, show that $\lambda = (\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e})/(\mathbf{e}^{\mathrm{T}}\mathbf{e})$. HINT: Recall that the dot product of two column vectors $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^{\mathrm{T}}\mathbf{v}$.

14. Show that if $\mathbf{A}$ is triangular, its eigenvalues are simply the diagonal elements of $\mathbf{A}$.

15. Show that if λ is an eigenvalue of $\mathbf{A}$, with a corresponding eigenvector $\mathbf{e}$, then λ^n is an eigenvalue of $\mathbf{A}^n$, with the same eigenvector $\mathbf{e}$, for any integer n. (Of course, if n is negative, $\mathbf{A}$ needs to be nonsingular if $\mathbf{A}^n$ is to exist in the first place.) HINT: Pre-multiply $\mathbf{Ae} = \lambda\mathbf{e}$ by $\mathbf{A}, \mathbf{A}^2, \ldots$.

16. Use the results stated in the preceding two exercises to determine the eigenvalues and eigenspaces of $\mathbf{A}^{10}$ for each of the following $\mathbf{A}$ matrices. Check your results by working out $\mathbf{A}^{10}$ and its eigenvalues and eigenvectors.

(a) $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & 2 \\ 0 & 0 & 2 \end{bmatrix}$

17. For the given $\mathbf{A}$ matrix, use computer software to determine its eigenvalues and eigenspaces. Then, use computer software to obtain $\mathbf{A}^5$ and to determine its eigenvalues and eigenspaces. Then, verify the result stated in Exercise 15, for this case.

(a) $\begin{bmatrix} 2 & 2 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{bmatrix}$

18. (*Similar matrices*) (a) Suppose that $\mathbf{Ax} = \mathbf{y}$, where $\mathbf{A}$ is square. Setting $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$ and $\mathbf{y} = \mathbf{Q}\widetilde{\mathbf{y}}$, where $\mathbf{Q}$ is invertible, show that

$$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}} = \widetilde{\mathbf{y}}, \tag{18.1}$$

where

$$\widetilde{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{AQ}. \tag{18.2}$$

Given any invertible matrix $\mathbf{Q}$, matrices $\mathbf{A}$ and $\tilde{\mathbf{A}}$ related by (18.2) are said to be **similar**.

(b) Show that if $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are similar, then they have the same characteristic polynomials and hence the same eigenvalues.

19. (*The characteristic polynomial*) Let us write the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ in the standard form

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$
$$\equiv (-1)^n[\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - \cdots + (-1)^n\beta_n] = 0. \tag{19.1}$$

If we denote the n roots (which need not be real) as $\lambda_1, \lambda_2, \ldots, \lambda_n$, numbering repeated roots separately, we may factor

$$\lambda^n - \beta_1\lambda^{n-1} + \cdots + (-1)^n\beta_n = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n). \tag{19.2}$$

Multiplying out the right-hand side of (19.2) and equating coefficients of like powers of λ, on both sides of the equation, yields the relations

$$\begin{aligned} \beta_1 &= \lambda_1 + \lambda_2 + \cdots + \lambda_n, \\ \beta_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_{n-1}\lambda_n, \\ \beta_3 &= \lambda_1\lambda_2\lambda_3 + \cdots + \lambda_{n-2}\lambda_{n-1}\lambda_n, \\ &\vdots \\ \beta_n &= \lambda_1\lambda_2\cdots\lambda_n. \end{aligned} \tag{19.3}$$

For example, if $n = 4$, then $\beta_2 = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_3\lambda_4$ and $\beta_3 = \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_3\lambda_4 + \lambda_2\lambda_3\lambda_4$.

Alternatively, one may expand the determinant in (19.1) directly, and identify the coefficients of the various powers of λ in terms of certain subdeterminants of $\mathbf{A}$, and hence determine the β_j's in terms of these subdeterminants. The result, we state without proof, is that β_j (for $j = 1, \ldots, n$) *is the sum of all the jth-order principal minors of* $\mathbf{A}$. (By a "principal minor" we mean the determinant of a submatrix of $\mathbf{A}$, whose main diagonal lies along that of $\mathbf{A}$.) Thus,

$$\begin{aligned} \beta_1 &= a_{11} + a_{22} + \cdots + a_{nn} \equiv \mathrm{tr}\mathbf{A}, \\ &\vdots \\ \beta_n &= \det\mathbf{A}, \end{aligned} \tag{19.4}$$

where the sum $a_{11} + a_{22} + \cdots + a_{nn}$ is called the **trace** of $\mathbf{A}$ and is denoted here as $\mathrm{tr}\mathbf{A}$.

(a) Verify equations (19.4) for $n = 2$ by expanding $\det(\mathbf{A} - \lambda\mathbf{I})$.

(b) Verify equations (19.4) for $n = 3$ by expanding $\det(\mathbf{A} - \lambda\mathbf{I})$; i.e., verify that

$$\beta_1 = a_{11} + a_{22} + a_{33},$$
$$\beta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix},$$
$$\beta_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

(c) Finally, comparing equations (19.3) with equations (19.4), show that

$$\begin{aligned} \lambda_1 + \lambda_2 + \cdots + \lambda_n &= \mathrm{tr}\mathbf{A}, \\ &\vdots \\ \lambda_1\lambda_2\cdots\lambda_n &= \det\mathbf{A}, \end{aligned} \tag{19.5}$$

and hence that $\mathbf{A}$ is singular if and only if at least one of its eigenvalues is zero.

20. Show that if two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvalues $\lambda_1, \ldots, \lambda_n$ and the same n LI eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, then it must be true that $\mathbf{A} = \mathbf{B}$.

21. Can an $n \times n$ matrix have more than n LI eigenvectors? Explain.

22. (*Markov matrices*) Recall the definition of a Markov matrix, and the fact every Markov matrix contains $\lambda = 1$ among its eigenvalues. You may use the result stated in Exercise 11, if you need it.

(a) Find the eigenvalues of

$$\mathbf{A} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

(b) Find one eigenvalue of

$$\mathbf{A} = \begin{bmatrix} 8 & 10 & 12 \\ 9 & 10 & 11 \\ 10 & 10 & 10 \end{bmatrix}.$$

(c) Find one eigenvalue of

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 2 & 1 \\ 0 & 2 & 2 \end{bmatrix}.$$

(d) Determine the eigenvalues and a basis for each eigenspace for the 20×20 matrix $\mathbf{A}$ having unity for each of its 400 elements.
(e) Determine the eigenvalues and a basis for each eigenspace for the 30×30 matrix $\mathbf{A}$ having 2 for each of its 900 elements.

23. In each case, use the same method as in Example 3 to find the general solution of the given system of coupled differential equations, if possible. If your solution falls short of being a general solution, explain why that happened. Primes denote d/dt.

(a) $x' = x + 2y$
$y' = 3x + 6y$

(b) $x' = x + y$
$y' = x + y$

(c) $x'' = 2x + y$
$y'' = 9x + 2y$

(d) $x'' = x + y$
$y'' = x + y$

(e) $x' = 4x + y$
$y' = -5x + z$
$z' = x - y - z$

(f) $x' = y - z$
$y' = -5x + 4y + z$
$z' = x - z$

(g) $x' = 2x - y$
$y' = y - z$
$z' = -x + y$

(h) $x'' = -x + y$
$y'' = -x + z$
$z'' = x - 2z$

24. (*Cayley–Hamilton theorem*) The **Cayley–Hamilton theorem** states that if the characteristic equation of any square matrix $\mathbf{A}$ is $\lambda^n + \alpha_1 \lambda^{n-1} + \cdots + \alpha_{n-1}\lambda + \alpha_n = 0$, then $\mathbf{A}^n + \alpha_1 \mathbf{A}^{n-1} + \cdots + \alpha_{n-1}\mathbf{A} + \alpha_n \mathbf{I} = \mathbf{0}$; i.e., $\mathbf{A}$ satisfies its characteristic equation.

(a) Prove this theorem for the general 2×2 case, $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

(b) If $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, show that $\mathbf{A}^2 - 4\mathbf{A} + 3\mathbf{I} = \mathbf{0}$ so that

$$\mathbf{A}^{-1} = \frac{4}{3}\mathbf{I} - \frac{1}{3}\mathbf{A} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

25. (*Generalized eigenvalue problem*) If $\mathbf{B} \neq \mathbf{I}$, then $\mathbf{Ax} = \lambda \mathbf{Bx}$ is called a **generalized eigenvalue problem**. It should be easy to see that in this case the characteristic equation is $\det(\mathbf{A} - \lambda\mathbf{B}) = 0$, and that the eigenvectors then follow as the nontrivial solutions of $(\mathbf{A} - \lambda\mathbf{B})\mathbf{x} = \mathbf{0}$. Find the eigenvalues and eigenspaces in each case.

(a) $\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(b) $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(d) $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

26. In seeking the eigenvalues and eigenvectors of a given matrix $\mathbf{A}$, is it permissible first to simplify $\mathbf{A}$ by means of some elementary row operations? (That is, are the eigenvalues and eigenvectors of $\mathbf{A}$ invariant with respect to elementary row operations?) Explain.

27. In each case, given the values of a, b, c, d in (32), use the eigenvalues to classify the singularity as an unstable node, stable node, saddle, stable focus, center, or unstable focus. If applicable, use the eigenvectors to give a labeled sketch of any stable and/or unstable manifolds.

(a) $a = 1, b = 1, c = 3, d = -1$
(b) $a = -1, b = -3, c = 1, d = 1$
(c) $a = 4, b = 1, c = 1, d = 4$
(d) $a = -3, b = 1, c = 1, d = -3$
(e) $a = 3, b = 1, c = -1, d = 3$
(f) $a = -2, b = -2, c = 2, d = -2$
(g) $a = 1, b = 2, c = 3, d = 4$
(h) $a = 5, b = 1, c = -8, d = 1$

11.3 Symmetric Matrices

In applications, symmetric matrices arise surprisingly often, and their symmetry leads to important results regarding their eigenvalues and eigenvectors. Thus, it is important to treat this case separately.

11.3.1. Eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$. We have the following three important results. We shall prove the third, but defer proofs for the first two to the exercises.

THEOREM 11.3.1 *Real Eigenvalues*
If $\mathbf{A}$ is symmetric ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$), then all of its eigenvalues are real.

THEOREM 11.3.2 *Dimension of Eigenspace*
If an eigenvalue λ of a symmetric matrix $\mathbf{A}$ is of multiplicity k, then the eigenspace corresponding to λ is of dimension k.

THEOREM 11.3.3 *Orthogonality of Eigenvectors*
If $\mathbf{A}$ is symmetric, then eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof: Let $\mathbf{e}_j$ and $\mathbf{e}_k$ be eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively. Thus,

$$\mathbf{A}\mathbf{e}_j = \lambda_j \mathbf{e}_j \quad \text{and} \quad \mathbf{A}\mathbf{e}_k = \lambda_k \mathbf{e}_k. \tag{1a,b}$$

Next, recall that the dot product of n-dimensional column vectors $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, and that $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ for any matrices $\mathbf{A}$ and $\mathbf{B}$ that are conformable for multiplication. Then, if we dot $\mathbf{e}_k$ into each side of (1a) and dot each side of (1b) into $\mathbf{e}_j$ [i.e., we "pre-dot" (1a) with $\mathbf{e}_k$ and "post-dot" (1b) with $\mathbf{e}_j$], we obtain

$$\begin{array}{rl|rl} \mathbf{e}_k \cdot (\mathbf{A}\mathbf{e}_j) &= \mathbf{e}_k \cdot (\lambda_j \mathbf{e}_j) & (\mathbf{A}\mathbf{e}_k) \cdot \mathbf{e}_j &= (\lambda_k \mathbf{e}_k) \cdot \mathbf{e}_j \\ \mathbf{e}_k^{\mathrm{T}} \mathbf{A}\mathbf{e}_j &= \lambda_j \mathbf{e}_k^{\mathrm{T}} \mathbf{e}_j & (\mathbf{A}\mathbf{e}_k)^{\mathrm{T}} \mathbf{e}_j &= \lambda_k \mathbf{e}_k^{\mathrm{T}} \mathbf{e}_j \\ & & \mathbf{e}_k^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{e}_j &= \lambda_k \mathbf{e}_k^{\mathrm{T}} \mathbf{e}_j. \end{array} \tag{2}$$

But $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ by assumption so if we subtract the bottom equations on the left and right of the vertical divider, we obtain

$$0 = (\lambda_j - \lambda_k)\mathbf{e}_k^{\mathrm{T}} \mathbf{e}_j. \tag{3}$$

Finally, $\lambda_j - \lambda_k \neq 0$ since λ_j and λ_k were assumed to be distinct so it follows from (3) that $\mathbf{e}_k^{\mathrm{T}} \mathbf{e}_j = 0$. Thus, $\mathbf{e}_k \cdot \mathbf{e}_j = 0$, as claimed. ■

As usual, be careful not to read converses into theorems. For instance, Theorem 11.3.1 says that if $\mathbf{A}$ is symmetric, then its eigenvalues are real. It does *not* say that the eigenvalues of $\mathbf{A}$ are real if and only if $\mathbf{A}$ is symmetric. For instance, in each of the four examples of Section 11.2 the λ's are real, yet none of the matrices is symmetric.

EXAMPLE 1. For the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \tag{4}$$

we find (Exercise 1) the eigenvalues and eigenspaces

$$\lambda_1 = 4, \ \mathbf{e}_1 = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \lambda_2 = 1, \ \mathbf{e}_2 = \beta \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \tag{5}$$

where $\lambda_2 = 1$ is of multiplicity two [i.e., the characteristic polynomial can be factored as $-(\lambda-4)(\lambda-1)^2$]. Since $\mathbf{A}$ is symmetric, the theorems apply. In accordance with Theorem 11.3.1, the λ's are real; in accordance with Theorem 11.3.2, λ_1 is of multiplicity 1 and its eigenspace is one-dimensional (namely, span$\{[1,1,1]^{\mathrm{T}}\}$), and λ_2 is of multiplicity 2 and its eigenspace is two-dimensional (namely, span$\{[-1,0,1]^{\mathrm{T}}, [-1,1,0]^{\mathrm{T}}\}$); and in accordance with Theorem 11.3.3, $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$ for all choices of α, β, γ. The eigenspace $\mathbf{e}_2$ is the plane through the origin (in 3-space) that is spanned by $[-1,0,1]^{\mathrm{T}}$ and $[-1,1,0]^{\mathrm{T}}$, and the eigenspace $\mathbf{e}_1$ is the line through the origin that is spanned by $[1,1,1]^{\mathrm{T}}$ and is normal to the plane.

The vectors $[-1,0,1]^{\mathrm{T}}$ and $[-1,1,0]^{\mathrm{T}}$ in $\mathbf{e}_2$ are LI and a basis for $\mathbf{e}_2$, but happen not to be orthogonal. Their lack of orthogonality does not violate Theorem 11.3.3 since they come from the same λ, not from distinct λ's. Nonetheless, we can "trade those vectors in" for two within $\mathbf{e}_2$ that *are* orthogonal (in a nonunique way, in fact, for there is an infinite number of pairs of orthogonal vectors within that plane). For instance, we can choose

$$\mathbf{e}_2 = [-1,0,1]^{\mathrm{T}} \tag{6}$$

[i.e., by setting $\beta = 1$ and $\gamma = 0$ in (5)] and seek

$$\mathbf{e}_3 = \beta[-1,0,1]^{\mathrm{T}} + \gamma[-1,1,0]^{\mathrm{T}} \tag{7}$$

such that

$$\mathbf{e}_2 \cdot \mathbf{e}_3 = (-1)(-\beta - \gamma) + (0)(\gamma) + (1)(\beta) = 2\beta + \gamma = 0. \tag{8}$$

Choosing $\beta = 1$, say, then $\gamma = -2$, and (7) gives

$$\mathbf{e}_3 = [1,-2,1]^{\mathrm{T}}, \tag{9}$$

and the vectors given by (6) and (9) constitute an orthogonal basis for the eigenspace corresponding to the eigenvalue 1.

And since (with $\alpha = 1$, say)

$$\mathbf{e}_1 = [1,1,1]^{\mathrm{T}} \tag{10}$$

is orthogonal to each of those vectors, it follows that the eigenvectors given by (6), (9), and (10) constitute an orthogonal basis for 3-space. That is, among the eigenvectors of the 3×3 symmetric $\mathbf{A}$ given by (4) we can find an orthogonal basis for 3-space.

COMMENT 1. The $\mathbf{A}$ matrix in (4) happens to be symmetric about the other diagonal as well as about the main diagonal. That symmetry is irrelevant; by symmetry we always mean that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, which is symmetry about the *main* diagonal (from upper left to bottom right).

COMMENT 2. You might be thinking "Of course the eigenvalues are real, for the $\mathbf{A}$ matrix is real." No, $\mathbf{A}$ being real implies only that the coefficients are real in its characteristic equation, and a polynomial equation with real coefficients *can* have complex roots. For instance, the real but nonsymmetric matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ has the characteristic equation $\lambda^2 - 2\lambda + 2 = 0$ and the complex eigenvalues $\lambda = 1 \pm i$.

COMMENT 3. The procedure that we used to obtain the orthogonal set $\{\mathbf{e}_2, \mathbf{e}_3\}$ from the LI set $\{[-1, 0, 1]^{\mathrm{T}}, [-1, 1, 0]^{\mathrm{T}}\}$ is essentially the Gram–Schmidt orthogonalization procedure explained in Exercise 11 of Section 9.9. ▮

Generalization of the ideas contained in Example 1 yields the following theorem.

THEOREM 11.3.4 *Orthogonal Basis*
If an $n \times n$ matrix $\mathbf{A}$ is symmetric, then its eigenvectors provide an orthogonal basis for n-space.

Proof: If all of $\mathbf{A}$'s n eigenvalues are distinct then, according to Theorem 11.3.3, the n eigenspaces are orthogonal (each being a one-dimensional line in n-space) and therefore provide n orthogonal vectors, which necessarily constitute a basis for n-space. What if the eigenvalues are not distinct? Suppose that all are distinct except for one, say λ, which is of multiplicity k. Then the $n - k$ eigenvectors corresponding to the other eigenvalues are orthogonal to each other and also to all vectors in the eigenspace corresponding to λ. Further, λ's eigenspace is k-dimensional (Theorem 11.3.2) and hence contains k orthogonal vectors. Altogether, then we have $(n - k) + k = n$ orthogonal eigenvectors and hence an orthogonal basis for n-space. A similar argument applies if there is more than one repeated eigenvalue. ■

EXAMPLE 2. *Free Vibration of a Two-Mass System.* Consider the system of two masses subjected to forces $f_1(t)$ and $f_2(t)$ and restrained laterally by springs and supported vertically by a frictionless table as shown in Fig. 1. The equations of motion, already derived in Example 3 of Section 3.9.1, are

$$\begin{aligned} m_1 x_1'' + (k_1 + k_{12})x_1 - k_{12}x_2 &= f_1(t), \\ m_2 x_2'' - k_{12}x_1 + (k_2 + k_{12})x_2 &= f_2(t). \end{aligned} \tag{11}$$

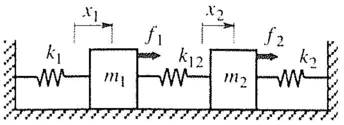


Figure 1. Two-mass system.

Let $m_1 = m_2 = k_1 = k_{12} = k_2 = 1$, say, for definiteness, and consider the *free vibration*, where $f_1(t) = f_2(t) = 0$. Then (11) becomes

$$\begin{aligned} x_1'' + 2x_1 - x_2 &= 0 \\ x_2'' - x_1 + 2x_2 &= 0. \end{aligned} \tag{12}$$

The system (12) is solved in Example 8 in Section 3.9.3, by the method of elimination, and we urge you to review that solution and to compare it with the following matrix eigenvalue problem approach. [We could also solve (12) by the Laplace transform method.]

Let us follow the same line of approach that was put forward in Example 2 of Section 11.2. Namely, seek

$$\begin{aligned} x_1(t) &= q_1 e^{\lambda t}, \\ x_2(t) &= q_2 e^{\lambda t}. \end{aligned} \tag{13}$$

Actually, on physical grounds we expect the solution to be a vibration (i.e., the λ's will turn out to be purely imaginary) so it seems more sensible to seek

$$\begin{aligned} x_1(t) &= q_1 \sin(\omega t + \phi), \\ x_2(t) &= q_2 \sin(\omega t + \phi), \end{aligned} \tag{14}$$

where the q_j's are amplitudes, ω is the frequency, and ϕ is the phase angle.* Putting (14) into (12) and canceling the $\sin(\omega t + \phi)$ factors gives

$$\begin{aligned} -\omega^2 q_1 + 2q_1 - q_2 = 0, \\ -\omega^2 q_2 - q_1 + 2q_2 = 0 \end{aligned}$$

or, equivalently,

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}, \tag{15}$$

which is a matrix eigenvalue problem

$$\mathbf{Aq} = \lambda \mathbf{q}, \tag{16}$$

with $\lambda = \omega^2$ as the eigenvalue. Solving for the eigenvalues and eigenspaces as explained in Section 11.2 we obtain

$$\lambda_1 = 1,\ \mathbf{e}_1 = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 3,\ \mathbf{e}_2 = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{17}$$

Each "eigenpair" gives us a solution of the form (14). The first gives $\omega = \sqrt{\lambda_1} = 1$, and†

$$\mathbf{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(t + \phi_1), \tag{18}$$

where α and ϕ_1 are arbitrary. The second gives $\omega = \sqrt{\lambda_2} = \sqrt{3}$, and

$$\mathbf{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\sqrt{3}\,t + \phi_2), \tag{19}$$

*Along the lines of COMMENT 1 in Example 2 of Section 11.2, if we do not use the same ω's and the same ϕ's in (14), then we will obtain only the trivial solution $x_1(t) = x_2(t) = 0$.

†The other root, $\omega = -1$, would yield no additional information.

where β and ϕ_2 are arbitrary, and there is no reason why ϕ_2 should be the same as ϕ_1. One can verify that (18) satisfies (12) for any α and any ϕ_1 and that (19) also satisfies (12) for any β and any ϕ_2. Since (12) is linear and homogeneous, it follows that the linear combination

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(t+\phi_1) + \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\sqrt{3}\,t+\phi_2) \tag{20}$$

is also a solution; indeed, it is a general solution of (12). Or, returning to scalar form, we have

$$\begin{aligned} x_1(t) &= \alpha \sin(t+\phi_1) + \beta \sin(\sqrt{3}\,t+\phi_2), \\ x_2(t) &= \alpha \sin(t+\phi_1) - \beta \sin(\sqrt{3}\,t+\phi_2). \end{aligned} \tag{21}$$

Here $\alpha, \beta, \phi_1, \phi_2$ are the constants of integration [just as A, B are the constants of integration in the general solution $x(t) = Ae^{3t} + Be^{-3t}$ of $x'' - 9x = 0$], and are determined from the initial conditions $x_1(0), x_2(0), x_1'(0), x_2'(0)$.

COMMENT 1. Note the central and organizing role of the eigenvalue problem. Each eigenpair defines a vibrational "mode," the eigenvalue gives the vibrational frequency ($\omega = \sqrt{\lambda}$) and the eigenvector gives the mode shape or configuration. The frequencies are called the **eigenfrequencies**, or **natural frequencies** (natural in that they correspond to the free, unforced vibration). The two terms on the right-hand side of (20) are called the **orthogonal modes** of vibration, orthogonal because $[1,1]^{\mathrm{T}} \cdot [1,-1]^{\mathrm{T}} = 0$, that orthogonality being a consequence of the symmetry of $\mathbf{A}$. The first term is called the **low mode** because it occurs at the lower of the two natural frequencies, and the second term is called the **high mode** because it is at the higher of those two frequencies.

COMMENT 2. Depending upon the initial conditions, we can excite either one of those modes or both of them. For instance, the conditions $x_1(0) = x_2(0) = 0$ and $x_1'(0) = x_2'(0) = 1$ give $\beta = 0, \alpha = 1, \phi_1 = 0$ (ϕ_2 is irrelevant because $\beta = 0$) and hence the low mode motion

$$\begin{aligned} x_1(t) &= \sin t, \\ x_2(t) &= \sin t, \end{aligned} \tag{22}$$

the conditions $x_1'(0) = x_2'(0) = 0$ and $x_1(0) = 1, x_2(0) = -1$ give $\alpha = 0, \beta = 1, \phi_2 = \pi/2$ (ϕ_1 is irrelevant because $\alpha = 0$) and hence the high mode motion

$$\begin{aligned} x_1(t) &= \sin(\sqrt{3}\,t + \frac{\pi}{2}) = \cos\sqrt{3}\,t, \\ x_2(t) &= -\sin(\sqrt{3}\,t + \frac{\pi}{2}) = -\cos\sqrt{3}\,t, \end{aligned} \tag{23}$$

and the conditions $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$ give a motion containing both modes,

$$\begin{aligned} x_1(t) &= \frac{1}{2}\sin(t+\frac{\pi}{2}) + \frac{1}{2}\sin(\sqrt{3}\,t+\frac{\pi}{2}) = \frac{1}{2}\cos t + \frac{1}{2}\cos\sqrt{3}\,t, \\ x_2(t) &= \frac{1}{2}\sin(t+\frac{\pi}{2}) - \frac{1}{2}\sin(\sqrt{3}\,t+\frac{\pi}{2}) = \frac{1}{2}\cos t - \frac{1}{2}\cos\sqrt{3}\,t. \end{aligned} \tag{24}$$

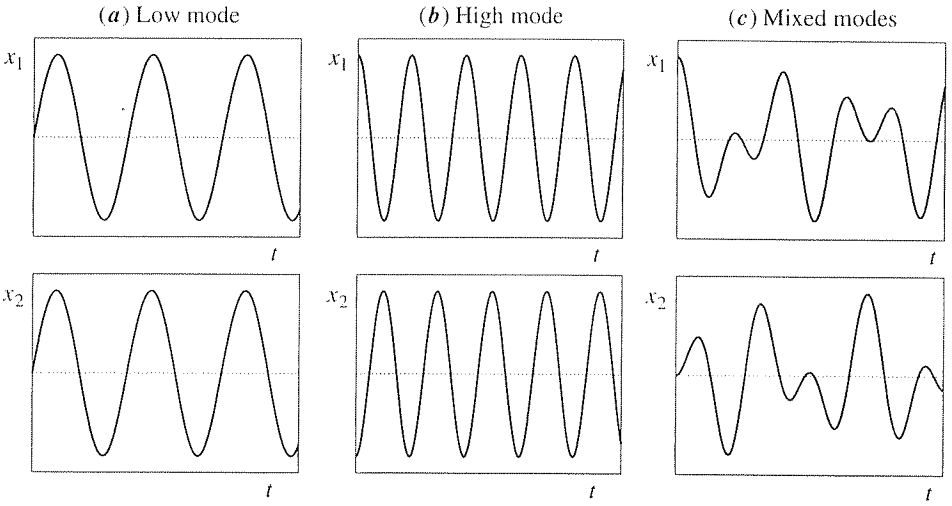


Figure 2. Low mode (22), high mode (23), mixed modes (24).

These three motions are contrasted in Fig. 2. Observe that each of the individual modes is simple and "clean," but the mixed mode motion is not. In the low mode the masses vibrate in unison and at the low frequency $\omega = 1$, and in the high mode their motions are opposite and at the high frequency $\omega = \sqrt{3}$.

To be sure it is clear how to apply the initial conditions, let us derive (24). We have the four equations

$$\begin{aligned} x_1(0) &= 1 = \alpha \sin \phi_1 + \beta \sin \phi_2, \\ x_2(0) &= 0 = \alpha \sin \phi_1 - \beta \sin \phi_2, \\ x_1'(0) &= 0 = \alpha \cos \phi_1 + \sqrt{3}\beta \cos \phi_2, \\ x_2'(0) &= 0 = \alpha \cos \phi_1 - \sqrt{3}\beta \cos \phi_2 \end{aligned} \tag{25}$$

in $\alpha, \beta, \phi_1, \phi_2$. They happen not to be linear algebraic equations [i.e., of the form $(\,)\alpha + (\,)\beta + (\,)\phi_1 + (\,)\phi_2 = (\,)$, where the parentheses contain constants], but they are readily solved. For instance, adding the first two and last two gives

$$\alpha \sin \phi_1 = \frac{1}{2}, \quad \alpha \cos \phi_1 = 0, \tag{26}$$

and these give $\phi_1 = \pi/2$ and $\alpha = 1/2$.* Similarly, subtracting the second from the first and the fourth from the third results in $\phi_2 = \pi/2$ and $\beta = 1/2$.

COMMENT 3. In this example there were two masses and two "degrees of freedom," $x_1(t)$ and $x_2(t)$. Consequently, **A** was 2×2 and the motion of each mass was found to consist of a linear combination of two eigenmodes. More generally, if there are n masses and n degrees of freedom $x_1(t), \ldots, x_n(t)$, then the motion of each mass consists of a linear

*These values are not uniquely determined. For instance $\phi_1 = 3\pi/2$ and $\alpha = -1/2$ satisfy (26) too. However, such differences do not lead to different solutions $x_1(t)$ and $x_2(t)$.

combination of n eigenmodes of an $n \times n$ matrix. ▮

11.3.2. Nonhomogeneous problem $\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}$. (Optional) In Chapters 8 and 10 we studied the general nonhomogeneous equation $\mathbf{Ax} = \mathbf{c}$, and in this chapter we have studied the eigenvalue problem, which is the homogeneous equation $\mathbf{Ax} = \lambda\mathbf{x}$. Next, we wish to show that eigenvalue problem concepts can be used to solve *non*homogeneous equations.

Specifically, consider the nonhomogeneous equation

$$\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}, \tag{27}$$

where the scalar Λ is a parameter.* We ask of $\mathbf{A}$ that it be $n \times n$, and that its eigenvectors provide a basis (not necessarily orthogonal) for $\mathbb{R}^n$. That will surely be the case if $\mathbf{A}$ is symmetric since then its eigenvectors provide an orthogonal basis for $\mathbb{R}^n$.

Of course, since Λ is considered as known we could subtract $\Lambda\mathbf{x}$ from $\mathbf{Ax}$ and absorb Λ into the $\mathbf{A}$ matrix. However, if Λ is a design parameter and we wish to see the explicit effect of Λ on the solution, then it is best to leave the $\Lambda\mathbf{x}$ term intact.

The idea is that to solve (27) we first take a "time-out" and solve the eigenvalue problem for $\mathbf{A}$ (i.e., $\mathbf{Ax} = \lambda\mathbf{x}$); that is, solve for the eigenvalues and eigenvectors of $\mathbf{A}$, which we denote as $\lambda_1, \dots, \lambda_n$ (not necessarily distinct) and $\mathbf{e}_1, \dots, \mathbf{e}_n$. Next, expand both $\mathbf{x}$ and $\mathbf{c}$, in (27), in terms of the $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ basis:

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{e}_j \quad \text{and} \quad \mathbf{c} = \sum_{j=1}^{n} c_j \mathbf{e}_j. \tag{28a,b}$$

The c_j's are known (i.e., they can be computed) since we know $\mathbf{c}$ and the $\mathbf{e}_j$ base vectors so the a_j's are our unknowns. To evaluate them, put (28) into (27):

$$\mathbf{A}\sum_{1}^{n} a_j \mathbf{e}_j = \Lambda \sum_{1}^{n} a_j \mathbf{e}_j + \sum_{1}^{n} c_j \mathbf{e}_j. \tag{29}$$

But

$$\mathbf{A}\sum_{1}^{n} a_j \mathbf{e}_j = \sum_{1}^{n} a_j \mathbf{A}\mathbf{e}_j = \sum_{1}^{n} a_j \lambda_j \mathbf{e}_j, \tag{30}$$

so that (29) can be re-expressed as

$$\sum_{1}^{n} (\lambda_j - \Lambda) a_j \mathbf{e}_j = \sum_{1}^{n} c_j \mathbf{e}_j. \tag{31}$$

Finally, since the $\mathbf{e}_j$'s are LI (for they are a basis) it follows from (31) that

$$(\lambda_j - \Lambda) a_j = c_j. \quad (j = 1, \dots, n) \tag{32}$$

At this point we need to be careful to distinguish these cases:

*By a *parameter* we mean a constant, the value of which we are free to specify. In the equation $x'' + 9x = F \sin \Omega t$, for instance, F and Ω are parameters.

(i) Suppose that none of the λ_j's equals Λ. Then we can divide both sides of (32) by $\lambda_j - \Lambda$ and obtain $a_j = c_j/(\lambda_j - \Lambda)$, and the *unique solution*

$$\mathbf{x} = \sum_1^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j \tag{33}$$

of (27).

(ii) Next, suppose that $\Lambda = \lambda_1$, where λ_1 is an eigenvalue of multiplicity 1. Then (32) becomes $(0)a_1 = c_1$ for $j = 1$, and two possibilities exist: if $c_1 \neq 0$, then there exists no a_1 satisfying $(0)a_1 = c_1$, and there is *no solution* of (27); but if $c_1 = 0$, then a_1 is arbitrary, and (27) admits a *nonunique* solution, the one-parameter family of solutions

$$\mathbf{x} = a_1\mathbf{e}_1 + \sum_2^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j. \quad (a_1 \text{ arbitrary}) \tag{34}$$

(iii) Similarly if $\Lambda = \lambda_1$ is an eigenvalue of multiplicity p: if $c_1, \dots, c_p$ are not all zero, then there is *no solution* of (27); and if $c_1 = \cdots = c_p = 0$, then there is a *nonunique* solution, the p-parameter family of solutions

$$\mathbf{x} = a_1\mathbf{e}_1 + \cdots + a_p\mathbf{e}_p + \sum_{p+1}^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j. \quad (a_1, \dots, a_p \text{ arbitrary}) \tag{35}$$

It is illuminating to compare (35) with the solution (18) in Section 10.5, to the problem $\mathbf{Ax} = \mathbf{c}$ (i.e., $\Lambda = 0$ in that case). There, $\mathbf{x}_0$ is a particular solution (i.e., $\mathbf{Ax}_0 = \mathbf{c}$), and $\mathbf{x}_1, \dots, \mathbf{x}_p$ were homogeneous solutions (i.e., $\mathbf{Ax}_1 = \cdots = \mathbf{Ax}_p = \mathbf{0}$). In (35), the $\sum_{p+1}^n$ term corresponds to $\mathbf{x}_0$, and $\mathbf{e}_1, \dots, \mathbf{e}_p$ are homogeneous solutions [i.e., solutions of (27) with $\mathbf{c}$ removed].[*]

EXAMPLE 3. *Forced Vibration of the Two-Mass System.* To illustrate, consider the mechanical system of Example 2 again, but this time with the forcing functions $f_1(t) = F_1 \sin \Omega t$ and $f_2(t) = F_2 \sin \Omega t$. Then in place of (12) we have

$$\begin{aligned} x_1'' + 2x_1 - x_2 &= F_1 \sin \Omega t, \\ x_2'' - x_1 + 2x_2 &= F_2 \sin \Omega t. \end{aligned} \tag{36}$$

We already found the general solution of the homogeneous equations (12), so in this example let us seek only a particular solution (the total solution then being the sum of the two). Let us seek a particular solution in the form[†]

$$\begin{aligned} x_1(t) &= q_1 \sin \Omega t, \\ x_2(t) &= q_2 \sin \Omega t. \end{aligned} \tag{37}$$

[*] Of course there is no reason why the $\mathbf{e}_1, \dots, \mathbf{e}_p$ need to equal $\mathbf{x}_1, \dots, \mathbf{x}_p$, but it *is* true that $\operatorname{span}\{\mathbf{e}_1, \dots, \mathbf{e}_p\}$ and $\operatorname{span}\{\mathbf{x}_1, \dots, \mathbf{x}_p\}$ are identical.

[†] The form (37) is inspired by the method of undetermined coefficients. Actually, we might try $x_1(t) = q_1 \sin \Omega t + r_1 \cos \Omega t$ and $x_2(t) = q_2 \sin \Omega t + r_2 \cos \Omega t$, but we can anticipate that the $\cos \Omega t$ terms will not be needed (i.e., we will find that $r_1 = r_2 = 0$) because there are no x_1' or x_2' terms on the left-hand side of (36).

Putting (37) into (36), canceling the $\sin \Omega t$ factors, and expressing the equations in matrix form, we have

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \Omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \tag{38}$$

or

$$\mathbf{Aq} = \Lambda \mathbf{q} + \mathbf{c}, \tag{39}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \Lambda = \Omega^2, \quad \mathbf{c} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}.$$

Recall from (17) that the eigenvalues and eigenvectors of $\mathbf{A}$ are (with $\alpha = \beta = 1$, say)

$$\lambda_1 = 1, \ \mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 3, \ \mathbf{e}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{40}$$

Suppose that $\Lambda \neq \lambda_1$ and $\Lambda \neq \lambda_2$; that is, the driving frequency Ω does not equal either of the natural frequencies 1 and $\sqrt{3}$. Then we have the unique solution (33), where "x" is $\mathbf{q}$, $c_1 = (\mathbf{c} \cdot \mathbf{e}_1)/(\mathbf{e}_1 \cdot \mathbf{e}_1) = (F_1 + F_2)/2$, $c_2 = (\mathbf{c} \cdot \mathbf{e}_2)/(\mathbf{e}_2 \cdot \mathbf{e}_2) = (F_1 - F_2)/2$, $\lambda_1 = 1$, $\lambda_2 = 3$, and $\Lambda = \Omega^2$.* Thus,

$$\begin{aligned} \mathbf{q} &= \frac{F_1 + F_2}{2\,(1 - \Omega^2)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{F_1 - F_2}{2\,(3 - \Omega^2)} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= \frac{1}{(1 - \Omega^2)\,(3 - \Omega^2)} \begin{bmatrix} \left(2 - \Omega^2\right) F_1 + F_2 \\ F_1 + \left(2 - \Omega^2\right) F_2 \end{bmatrix} \end{aligned} \tag{41}$$

so the desired particular solution (37), of (36), is

$$x_1(t) = \frac{\left(2 - \Omega^2\right) F_1 + F_2}{(1 - \Omega^2)\,(3 - \Omega^2)} \sin \Omega t, \tag{42a}$$

$$x_2(t) = \frac{F_1 + \left(2 - \Omega^2\right) F_2}{(1 - \Omega^2)\,(3 - \Omega^2)} \sin \Omega t. \tag{42b}$$

In Section 3.8 we studied the forced vibration of a single mass, and stressed the importance to an engineer of the amplitude- and phase-response curves. Here too, let us plot the amplitude-response curves (Fig. 3) for the representative case where $F_1 = 1$ and $F_2 = 0$. The amplitudes $A_1(\Omega)$ and $A_2(\Omega)$ are, from (42a) and (42b),

$$A_1(\Omega) = \left| \frac{\left(2 - \Omega^2\right)}{(1 - \Omega^2)\,(3 - \Omega^2)} \right| \quad \text{and} \quad A_2(\Omega) = \left| \frac{1}{(1 - \Omega^2)\,(3 - \Omega^2)} \right|, \tag{43}$$

and we observe that these tend to infinity as Ω tends to either of the natural frequencies, 1 and $\sqrt{3}$.

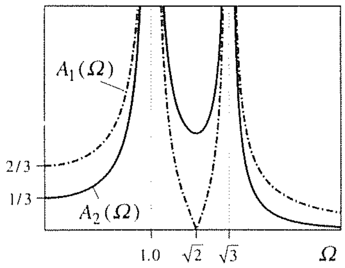


Figure 3. Amplitude-response curves, equation (43).

What if Ω *equals* 1 or $\sqrt{3}$? Then our derivation of (43) does not hold since it is based on the assumption that $\Lambda \neq \lambda_1$ and $\Lambda \neq \lambda_2$. If we do have $\Lambda = \lambda_1$ (i.e., $\Omega = 1$) say,

*If the formula $c_j = (\mathbf{c} \cdot \mathbf{e}_j)/(\mathbf{e}_j \cdot \mathbf{e}_j)$ is unfamiliar to you, we urge you to review Section 9.9, especially equation (23) therein.

then [according to case (ii) above] there is no solution if $c_1 \neq 0$ and there is an infinity of solutions if $c_1 = 0$. Since

$$c_1 = \frac{[F_1, F_2]^{\mathrm{T}} \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \tag{44}$$

the idea is that there will be no solution unless the forcing vector $[F_1, F_2]^{\mathrm{T}}$ does not excite that particular eigenmode – namely, unless $[F_1, F_2]^{\mathrm{T}}$ is orthogonal to $\mathbf{e}_1$! Similarly, if $\Lambda = \lambda_2$ (i.e., $\Omega = \sqrt{3}$).

What do we mean when we say there is no solution if $\Lambda = \lambda_1$ and $c_1 \neq 0$ (or $\Lambda = \lambda_2$ and $c_2 \neq 0$)? We do not mean that there exists no particular solution of (36) but only that there is no solution of the assumed form (37). A modified version of (37) is needed, but we will not pursue that point. ∎

The method presented here for solving (27) is known as the **eigenvector expansion method**. What is the advantage in using the eigenvectors of $\mathbf{A}$ as our basis; why not use *any* basis for $\mathbb{R}^n$? The idea is that the vector $\mathbf{Ae}_j$ in the middle of (30) is simply $\lambda_j \mathbf{e}_j$ if $\mathbf{e}_j$ is an eigenvector of $\mathbf{A}$, whereas it would be a linear combination of all the base vectors if some other basis were used. In that case we would end up with a *coupled* system of linear algebraic equations for the a_j's, rather than the uncoupled (and hence readily solved) system (32). This same comment applies to Example 4 in Section 11.2 where, to study the stability of the equilibrium population, we expanded the initial population vector $\mathbf{p}_0$ in terms of the eigenvectors of $\mathbf{A}$.

Closure. Symmetric matrices arise frequently in applications, and their symmetry leads to several important results regarding their eigenvalues and eigenvectors: their eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and the eigenvectors of an $n \times n$ symmetric matrix provide an orthogonal basis for $\mathbb{R}^n$.

We discussed an important application to multimass mechanical systems, and found that the free oscillation can be represented as the superposition of orthogonal modes, with the eigenvalues giving the modal frequencies and the eigenvectors giving the modal configurations.

The special importance of symmetric matrices is further revealed in the remaining sections of this chapter.

In the second half of this section we return to the nonhomogeneous equation $\mathbf{Ax} = \mathbf{c}$, actually to the form $\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}$ where, Λ is a parameter, and develop a line of approach known as the eigenvector expansion method. The idea behind that method is to compute first the eigenvalues and eigenvectors of $\mathbf{A}$. Assuming that we can obtain a basis of $\mathbb{R}^n$ from the eigenvectors of $\mathbf{A}$ (as is always possible for symmetric $\mathbf{A}$'s which, indeed, provide us with *orthogonal* bases), we use that basis, which is the most convenient or "natural" basis to expand $\mathbf{x}$ and $\mathbf{c}$. Finally, equating coefficients of the various base vectors on both sides of the equation gives uncoupled linear algebraic equations on the unknown coefficients in the expansion of $\mathbf{x}$.

The eigenvector expansion method is applicable to other cases as well, such as ordinary and partial differential equations. We will meet it again when we come to partial differential equations.

EXERCISES 11.3

1. From the eigenvectors of the given $n \times n$ matrix obtain an orthogonal basis for $\mathbb{R}^n$.

(a) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix}$

(e) $\begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$ (f) $\begin{bmatrix} 7 & 4 & -4 \\ 4 & 1 & 8 \\ -4 & 8 & 1 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & 4 & 0 \\ 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 3 & 3 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}$

(i) $\begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

2. Determine whether or not the eigenvectors of the given *non*symmetric 2×2 matrix provide an orthogonal basis for $\mathbb{R}^2$.

(a) $\begin{bmatrix} 0 & 0 \\ 3 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 9 \\ 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 \\ 9 & 0 \end{bmatrix}$

3. (*Proof of Theorem 11.3.1*) (a) Prove Theorem 11.3.1 for the simple case where $\mathbf{A}$ is merely 2×2, i.e., of the form $\begin{bmatrix} a & b \\ b & d \end{bmatrix}$.

(b) We will now supply the "skeleton" of a general proof of Theorem 11.3.1 and ask you to answer questions about the steps. Overhead bars will denote complex conjugates: if $z = x + iy$, then $\overline{z} = x - iy$. Proof:

$$\mathbf{Ae} = \lambda \mathbf{e}, \tag{3.1}$$

so

$$\mathbf{A}\overline{\mathbf{e}} = \overline{\lambda}\overline{\mathbf{e}}. \tag{3.2}$$

Dot both sides of (3.1) into $\overline{\mathbf{e}}$, and dot $\mathbf{e}$ into both sides of (3.2):

$$(\mathbf{Ae}) \cdot \overline{\mathbf{e}} = (\lambda \mathbf{e}) \cdot \overline{\mathbf{e}} \quad \text{and} \quad \mathbf{e} \cdot (\mathbf{A}\overline{\mathbf{e}}) = \mathbf{e} \cdot (\overline{\lambda}\overline{\mathbf{e}}). \tag{3.3}$$

Thus,

$$\mathbf{e}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\overline{\mathbf{e}} = \lambda \mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}} \quad \text{and} \quad \mathbf{e}^{\mathrm{T}}\mathbf{A}\overline{\mathbf{e}} = \overline{\lambda}\mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}}, \tag{3.4}$$

so

$$(\lambda - \overline{\lambda})\mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}} = 0, \tag{3.5}$$

and hence

$$\lambda = \overline{\lambda} \tag{3.6}$$

so λ is real. Questions: How does (3.2) follow from (3.1)? (3.4) from (3.3)? (3.5) from (3.4)? (3.6) from (3.5) ?

4. (*Proof of Theorem 11.3.2*) Prove Theorem 11.3.2 for the simple case where $n = 2$.

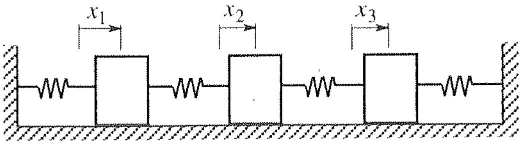
5. (*Post-dotting versus pre-dotting*) Observe that the top left equation in (2) was obtained by dotting $\mathbf{e}_k$ into both sides of (1a); i.e., we "pre-dotted" (1a) with $\mathbf{e}_k$. However, the top right equation in (2) was obtained by dotting both sides of (1b) into $\mathbf{e}_j$; i.e., we "post-dotted" $\mathbf{e}_j$ into (1b). Now, post-dotting or pre-dotting doesn't matter, in the sense that the dot product is commutative: $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$. Nevertheless, one may be more convenient than the other. Specifically, show that if we write the top left equation in (2) as $(\mathbf{Ae}_j) \cdot \mathbf{e}_k = (\lambda_j \mathbf{e}_j) \cdot \mathbf{e}_k$, instead of $\mathbf{e}_k \cdot (\mathbf{Ae}_j) = \mathbf{e}_k \cdot (\lambda_j \mathbf{e}_j)$, then it is more difficult to obtain (3) and hence the desired result.

6. Use any theorem(s) from Chapter 3 to show that the solution to (12), with the initial conditions $x_1(0), x_2(0), x_1'(0)$, and $x_2'(0)$ specified, is unique.

7. Beginning with (21), complete the solution for the following initial conditions:

(a) $x_1(0) = 2, x_2(0) = 3, x_1'(0) = x_2'(0) = 0$
(b) $x_1(0) = 1, x_2(0) = x_1'(0) = 0, x_2'(0) = -3$
(c) $x_1(0) = x_2(0) = x_1'(0) = 0, x_2'(0) = 5$
(d) $x_1(0) = -2, x_2(0) = x_1'(0) = 0, x_2'(0) = 3$

8. Consider the three-mass system shown.



(a) Derive the equations of motion for the free vibration, taking all masses and spring stiffnesses to be 1, say.
(b) Following the same lines as in Example 2, find the orthogonal modes (i.e., the eigenvectors) and their corresponding natural frequencies, proceeding either by hand or by using computer software.
(c) Give any set of initial conditions that will excite the low frequency mode only; the middle mode only; the high mode only.
(d) Find the solution $x_1(t), x_2(t), x_3(t)$ corresponding to the initial condition $x_1(0) = 1, x_2(0) = x_3(0) = x_1'(0) = x_2'(0) = x_3'(0) = 0$.

9. Consider a mass-spring system like the one shown in Exercise 8, but with five masses and six springs. If all the masses and spring stiffnesses are 1, say, then the equations of motion are these, in matrix form:

$$\begin{bmatrix} x_1'' \\ x_2'' \\ x_3'' \\ x_4'' \\ x_5'' \end{bmatrix} + \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{9.1}$$

or $\mathbf{x}'' + \mathbf{A}\mathbf{x} = \mathbf{0}$. Observe that the $\mathbf{A}$ matrix is **tridiagonal** (i.e., all elements are zero except for the main diagonal and the two neighboring diagonals). Physically, this result correponds to **nearest-neighbor coupling** whereby each mass feels only its immediate neighbors. Nearest-neighbor coupling occurs in other systems as well. For instance, in modeling single-lane traffic flow, each driver accelerates or decelerates according to the motion of the cars immediately ahead and immediately behind, so the resulting coupled differential equations exhibit nearest-neighbor coupling. The problem that we pose is for you to use computer software to determine the natural frequencies and corresponding mode shapes (eigenvectors).

10. (*Beats*) Suppose in Example 2 that $m_1 = m_2 = k_{12} = 1$, $k_1 = k_2 = 20$ so that the coupling of m_1 and m_2 is weak since k_{12} is much smaller than k_1 and k_2. (In the limit, if k_{12} were zero, there would be no coupling at all, and the motions of m_1 and m_2 would be entirely independent.) Supposing further that $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$, proceed as in Example 2 and show that, for the free vibration,

$$\begin{aligned} x_1(t) &= \tfrac{1}{2}\left(\cos\sqrt{20}\,t + \cos\sqrt{22}\,t\right), \\ x_2(t) &= \tfrac{1}{2}\left(\cos\sqrt{20}\,t - \cos\sqrt{22}\,t\right). \end{aligned} \tag{10.1}$$

Next, use the trigonometric identities

$$\begin{aligned} \cos A + \cos B &= 2\cos\tfrac{A+B}{2}\cos\tfrac{A-B}{2}, \\ \cos A - \cos B &= -2\sin\tfrac{A+B}{2}\sin\tfrac{A-B}{2} \end{aligned} \tag{10.2}$$

to show that

$$\begin{aligned} x_1(t) &\approx \cos 4.58t \cos 0.11t, \\ x_2(t) &\approx \sin 4.58t \sin 0.11t. \end{aligned} \tag{10.3}$$

Use (10.3) to sketch $x_1(t)$ and $x_2(t)$ versus t in separate graphs, one below the other, labeling key values. Observe the slow transfer of energy back and forth between m_1 and m_2. In vibration theory this phenomenon is known as **beats**.

11. (*Rayleigh's quotient*) Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Dotting any eigenvector $\mathbf{e}$ of $\mathbf{A}$ into both sides of $\mathbf{Ae} = \lambda\mathbf{e}$ and solving for λ, gives

$$\lambda = \frac{\mathbf{e}\cdot\mathbf{Ae}}{\mathbf{e}\cdot\mathbf{e}} = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{Ae}}{\mathbf{e}^{\mathrm{T}}\mathbf{e}}. \tag{11.1}$$

More generally, if $\mathbf{x}$ is any vector, not necessarily an eigenvector of $\mathbf{A}$, then the number

$$\boxed{R(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{Ax}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}} \tag{11.2}$$

is known as **Rayleigh's quotient**, after *Lord Rayleigh* (*John William Strutt*; 1842–1919). Imagine putting randomly chosen $\mathbf{x}$ vectors, one after another, into Rayleigh's quotient. If $\mathbf{x}$ happens to coincide with an eigenvector, then, according to (11.1), $R(\mathbf{x})$ gives the corresponding eigenvalue. In any case,

$$|R(\mathbf{x})| = \left|\frac{\mathbf{x}^{\mathrm{T}}\mathbf{Ax}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}\right| \leq |\lambda_1|, \tag{11.3}$$

where the eigenvalues are ordered so that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. That is, $|R(\mathbf{x})|$ provides a **lower bound** on the magnitude of the largest eigenvalue of $\mathbf{A}$, where $\mathbf{x}$ is any vector [i.e., any nonzero vector, since $R(\mathbf{0}) = 0/0$ is undefined]. Upper and lower bounds on eigenvalues are sometimes important, and Rayleigh's quotient is used again in Exercise 12.

(a) Prove the inequality in (11.3). HINT: Since $\mathbf{A}$ is symmetric, it has n orthogonal eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$. Expand $\mathbf{x}$ in terms of the orthogonal basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$:

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{e}_j. \tag{11.4}$$

(b) Verify (11.3) for the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & -2 \\ 0 & -2 & -1 \end{bmatrix} \tag{11.5}$$

by taking $\mathbf{x} = [1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}, [0,0,1]^{\mathrm{T}}, [1,2,3]^{\mathrm{T}}$ and $[3,-1,4]^{\mathrm{T}}$, say, verifying that $|R(\mathbf{x})| \le |\lambda_1|$ in each case.

(c) Evaluate $R(\mathbf{x})$ for $\mathbf{x} = [0.4, 0.3, 0.3]^{\mathrm{T}}$, $[0.6, 0.2, 0.2]^{\mathrm{T}}$, $[0.8, 0.1, 0.1]^{\mathrm{T}}$, $[0.96, 0.02, 0.02]^{\mathrm{T}}$, $[1,0,0]^{\mathrm{T}}$, and for $\mathbf{x} = [0, 1, 1.4]^{\mathrm{T}}$, $[0,1,1.1]^{\mathrm{T}}$, $[0,1,1]^{\mathrm{T}}$, and discuss your results in the light of (11.1) and (11.3).

12. (*The power method*) There exists a simple *iterative* procedure for calculating eigenvalues and eigenvectors, which is known as the **power method**. To begin, one selects any nonzero vector $\mathbf{x}^{(0)}$ and then computes $\mathbf{x}^{(1)} \equiv \mathbf{A}\mathbf{x}^{(0)}$, $\mathbf{x}^{(2)} \equiv \mathbf{A}\mathbf{x}^{(1)}$, and so on. That is,

$$\mathbf{x}^{(k+1)} \equiv \mathbf{A}\mathbf{x}^{(k)} \quad (k = 0, 1, 2, \dots). \tag{12.1}$$

Before analyzing the situation, let us apply (12.1) and see what happens. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \tag{12.2}$$

for example. Choosing $\mathbf{x}^{(0)} = [1,0,0]^{\mathrm{T}}$, say, successive application of (12.1) gives

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \overset{\mathbf{x}^{(0)}}{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}} \overset{\mathbf{x}^{(1)}}{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}} \overset{\mathbf{x}^{(2)}}{\begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix}} \overset{\mathbf{x}^{(3)}}{\begin{bmatrix} 5 \\ 3 \\ 3 \end{bmatrix}} \overset{\mathbf{x}^{(4)}}{\begin{bmatrix} 11 \\ 5 \\ 5 \end{bmatrix}} \overset{\mathbf{x}^{(5)}}{\begin{bmatrix} 21 \\ 11 \\ 11 \end{bmatrix}}, \tag{12.3}$$

and so on. Now, observe, from the very nature of the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, that if $\mathbf{x}^{(0)}$ were an eigenvector of A, then $\mathbf{A}\mathbf{x}^{(0)}$ would be some scalar multiple of $\mathbf{x}^{(0)}$. But $\mathbf{A}\mathbf{x}^{(0)} \equiv \mathbf{x}^{(1)}$, and $\mathbf{x}^{(1)}$ is seen from (12.3) *not* to be a scalar multiple of $\mathbf{x}^{(0)}$. Hence, $\mathbf{x}^{(0)}$ is not an eigenvector of $\mathbf{A}$. Similarly, $\mathbf{x}^{(1)}$ is not an eigenvector because $\mathbf{x}^{(2)}$ is not a multiple of $\mathbf{x}^{(1)}$, and so on. Nevertheless, we see that with each successive step $\mathbf{x}^{(k+1)}$ draws closer and closer to being a multiple of $\mathbf{x}^{(k)}$ so that the sequence $\mathbf{x}^{(k)}$ is evidently *approaching* an eigenvector of $\mathbf{A}$. In fact, $\mathbf{x}^{(5)}$ is very close to being a multiple of $\mathbf{x}^{(4)}$ so that there is an eigenvector

$$\mathbf{e} \approx \begin{bmatrix} 21 \\ 11 \\ 11 \end{bmatrix}. \tag{12.4}$$

What is the corresponding λ, $\lambda \approx 21/11$? $\lambda \approx 11/5$? An average of the two? We state without proof that one does well to use the Rayleigh quotient from Exercise 11:

$$\lambda \approx \frac{\mathbf{x}^{(4)\mathrm{T}}\mathbf{A}\mathbf{x}^{(4)}}{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(4)}} = \frac{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(5)}}{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(4)}} = \frac{341}{171} = 1.994. \tag{12.5}$$

How do (12.4) and (12.5) compare with exact values? The eigenvalues and eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 2,\ \mathbf{e}_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \qquad \lambda_2 = -1,\ \mathbf{e}_2 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}, \qquad \lambda_3 = 0,\ \mathbf{e}_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \tag{12.6}$$

so the iteration (12.3) is evidently converging to $\mathbf{e}_1$. [It is striking that whereas (12.4) is accurate to only around one part in 20, (12.5) is accurate to around one part in 200. This enhancement of accuracy, using Rayleigh's quotient to determine λ, is not a coincidence and can be explained theoretically. See, for instance, Stephen H. Crandall, *Engineering Analysis* (New York: McGraw–Hill, 1956), Chap. 2.]

To see what is going on, suppose that $\mathbf{A}$ is a symmetric matrix of order n (although symmetry is more than we need; it would suffice for $\mathbf{A}$ to have n LI eigenvectors) and let its eigenvalues be ordered so that $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$ as in (12.6). Since $\mathbf{A}$ is symmetric, its eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ are on orthogonal basis for n-space. Hence, our initial vector $\mathbf{x}^{(0)}$ must be expressible as

$$\mathbf{x}^{(0)} = \sum_{j=1}^{n} a_j \mathbf{e}_j. \tag{12.7}$$

Of course, we cannot *compute* the a_j's since we do not know the $\mathbf{e}_j$'s yet, but that is no problem; it is the *form* of (12.7) that is important here. It follows from (12.1) and (12.7) that

$$\begin{aligned} \mathbf{x}^{(k)} &= \sum_{1}^{n} a_j \lambda_j^k \mathbf{e}_j \\ &= \lambda_1^k \left[a_1 \mathbf{e}_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{e}_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{e}_n \right]. \end{aligned}$$

(12.8)

If λ_1 is, in fact, *dominant*, i.e., $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$, then $(\lambda_2/\lambda_1)^k, \ldots, (\lambda_n/\lambda_1)^k$ all tend to zero as $k \to \infty$, so that $\mathbf{x}^{(k)} \sim \lambda_1^k a_1 \mathbf{e}_1$ as $k \to \infty$ (provided that $a_1 \neq 0$, i.e., provided that $\mathbf{x}^{(0)}$ does not happen to be orthogonal to $\mathbf{e}_1$). Eigenvectors can be scaled arbitrarily so the $\lambda_1^k a_1$ factor is of little interest; the point is that $\mathbf{x}^{(k)}$ converges to $\mathbf{e}_1$, the eigenvector corresponding to the dominant eigenvalue. That is precisely what was found in the preceding illustration, wherein the dominant eigenvalue is $\lambda_1 = 2$. NOTE: Once again we see that the most convenient basis to use is the basis provided by the **A** matrix itself.

(a) Show that (12.8) follows from (12.1) and (12.7).

(b) Determine the dominant eigenvalue and corresponding eigenvector by the power method for

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Also, evaluate the eigenvalues and eigenvectors exactly, either by hand or by using computer software. NOTE: Observe that even though your iteration converges, you cannot be certain that the eigenvalue obtained is the *dominant* one, λ_1, since your chosen $\mathbf{x}^{(0)}$ may (without your knowing it) be orthogonal to $\mathbf{e}_1$, as mentioned in the sentence following (12.8). Hence, we recommend that you carry out the iteration three times, once with $\mathbf{x}^{(0)} = [1, 0, 0]^{\mathrm{T}}$, once with $\mathbf{x}^{(0)} = [0, 1, 0]^{\mathrm{T}}$, and once with $\mathbf{x}^{(0)} = [0, 0, 1]^{\mathrm{T}}$ since there is no way that all *three* of these $\mathbf{x}^{(0)}$'s can be orthogonal to $\mathbf{e}_1$. (Do you see why this is so?) Go as far as $\mathbf{x}^{(4)}$ in each case, and use the Rayleigh quotient to estimate λ, as we did in (12.5).

13. The same as Exercise 12(b), for the given matrix

(a) $\begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 1 & -1 \\ 1 & 3 & 4 \\ -1 & 4 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 4 & 1 & 3 \\ 1 & 0 & -1 \\ 3 & -1 & 4 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 1 & 2 \end{bmatrix}$ (h) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

14. For the **A** matrix given in part (b) of Exercise 13, work out $\mathbf{A}^4$. Apply the power method (Exercise 13) both to **A** and $\mathbf{A}^4$, beginning with $\mathbf{x}^{(0)} = [0, 1, 0]^{\mathrm{T}}$, say. Go as far as $\mathbf{x}^{(2)}$ and use the Rayleigh quotient to estimate λ. Explain why the iteration converges more rapidly for $\mathbf{A}^4$ than for **A**, and show how to recover the eigenvalue and eigenvector of **A** from those of $\mathbf{A}^4$. (See Exercise 15 of Section 11.2.)

15. Consider the problem $\mathbf{A}\mathbf{x} = \Lambda\mathbf{x} + \mathbf{c}$, where $\mathbf{A}, \Lambda, \mathbf{c}$ are given below (along with the eigenvalues of **A**, for your convenience). Solve for **x** by the eigenfunction expansion method; if no solution exists, state that. You may use any of equations (33)–(35) without deriving them.

(a) **A** given in Exercise 13(a) ($\lambda = 7, 3, 0$), $\Lambda = 2$, $\mathbf{c} = [1, 2, 3]^{\mathrm{T}}$

(b) **A** given in Exercise 13(a) ($\lambda = 7, 3, 0$), $\Lambda = 3$, $\mathbf{c} = [2, 2, 0]^{\mathrm{T}}$

(c) **A** given in Exercise 13(a) ($\lambda = 7, 3, 0$), $\Lambda = 3$, $\mathbf{c} = [1, 1, 3]^{\mathrm{T}}$

(d) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 1$, $\mathbf{c} = [3, -1, 1, 0]^{\mathrm{T}}$

(e) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 4$, $\mathbf{c} = [1, 2, 0, 3]^{\mathrm{T}}$

(f) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 4$, $\mathbf{c} = [2, 0, 1, -2]^{\mathrm{T}}$

(g) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 0$, $\mathbf{c} = [1, 3, 3, 1]^{\mathrm{T}}$

(h) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 0$, $\mathbf{c} = [1, 2, 3, 4]^{\mathrm{T}}$

(i) **A** given in Exercise 13(g) ($\lambda = 4, 2, 0, 0$), $\Lambda = 0$, $\mathbf{c} = [1, 3, 3, 2]^{\mathrm{T}}$

16. To solve the nonhomogeneous equation (27), we first solve for the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of **A**, then we use those eigenvectors as a basis to expand **x** and **c**. Why do we go to the extra trouble of solving for the eigenvalues and eigenvectors of **A** when it is easy to make up bases [such as $(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$] without trouble? That is, explain what advantage there is in using the eigenvectors of **A** as our basis.

17. (*Generalized eigenvalue problem*) With $f_1(t) = f_2(t) = 0$, $x_1(t) = q_1 \sin(\omega t + \phi)$, and $x_2(t) = q_2 \sin(\omega t + \phi)$, (11) becomes the eigenvalue problem

$$\begin{bmatrix} \dfrac{k_1 + k_{12}}{m_1} & -\dfrac{k_{12}}{m_1} \\ -\dfrac{k_{12}}{m_2} & \dfrac{k_2 + k_{12}}{m_2} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \qquad (17.1)$$

or $\mathbf{A}\mathbf{q} = \lambda\mathbf{q}$, where $\lambda = \omega^2$. In Example 2 we took $m_1 = m_2 = 1$ so **A** is symmetric and the eigenmodes are orthogonal. In general, however, $m_1 \neq m_2$ and **A** is not symmetric. Nonetheless, observe that we can reexpress (16.1) as

$$\begin{bmatrix} k_1 + k_{12} & -k_{12} \\ -k_{12} & k_2 + k_{12} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \tag{17.2}$$

or

$$\mathbf{Kq} = \lambda \mathbf{Mq}, \tag{17.3}$$

where $\lambda = \omega^2$ again. Equation (17.3) is called a **generalized eigenvalue problem**, "generalized" because of the presence of $\mathbf{M}$. (The generalized eigenvalue problem is introduced in Exercise 20 of Section 11.2.) The eigenvalues are determined by $\det(\mathbf{K} - \lambda\mathbf{M}) = 0$, and then the corresponding eigenvectors follow as the nontrivial solutions of $(\mathbf{K} - \lambda\mathbf{M})\mathbf{q} = \mathbf{0}$. Here is the problem:

(a) Let $\mathbf{K}$ and $\mathbf{M}$ be symmetric [as is the case in (17.2), although there is no need for $\mathbf{M}$ to be diagonal as well]. Show that if $\mathbf{e}_j$ and $\mathbf{e}_k$ are eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively, then $\mathbf{e}_j$ and $\mathbf{e}_k$ satisfy the **generalized orthogonality relation**

$$\mathbf{e}_j \cdot (\mathbf{M}\mathbf{e}_k) = 0. \tag{17.4}$$

That is, $\mathbf{e}_j$ is orthogonal to $\mathbf{e}_k$ "relative to $\mathbf{M}$."
(b) Verify the truth of (17.4) for the case

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \lambda \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

11.4 Diagonalization

We have seen that diagonal matrices are particularly straightforward. For instance, the solution of $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is $n \times n$, is generally tedious if n is large but is simple if $\mathbf{A}$ is diagonal, for then the scalar equations are not coupled. Similarly, raising $\mathbf{A}$ to the mth power is generally tedious if m is large but is simple if $\mathbf{A}$ is diagonal. Likewise, the solution of a system of differential equations

$$\mathbf{x}'(t) = \mathbf{Ax}(t) \tag{1}$$

is generally tedious but is simple if $\mathbf{A}$ is diagonal, for then the scalar equations are uncoupled.

To introduce the idea of diagonalization, let us focus on the application of diagonalization to the solution of the system of differential equations given by (1), where we assume that $\mathbf{A}$ is constant (i.e., its elements do not vary with t). We have already studied several methods for the solution of (1): the method of elimination (Section 3.9), which is essentially Gauss elimination but where coefficients are differential operators; the Laplace transform method (Section 5.4), which would reduce (1) to n linear algebraic equations in $\bar{x}_1(s), \dots, \bar{x}_n(s)$; and seeking $\mathbf{x}(t) = \mathbf{q}e^{\lambda t}$ and obtaining an eigenvalue problem (Sections 11.2 and 11.3).

We begin the solution of (1) by diagonalization by making a linear change of variables from $x_1, \dots, x_n$ to $\tilde{x}_1, \dots, \tilde{x}_n$:

$$\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}, \tag{2}$$

where $\mathbf{Q}$ is a constant matrix. Written out,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{bmatrix}. \tag{3}$$

Putting (2) into (1) [and observing that $(\mathbf{Q}\tilde{\mathbf{x}})' = \mathbf{Q}\tilde{\mathbf{x}}'$ because the matrix $\mathbf{Q}$ is constant] gives

$$\mathbf{Q}\tilde{\mathbf{x}}' = \mathbf{A}\mathbf{Q}\tilde{\mathbf{x}}. \tag{4}$$

Since the choice of $\mathbf{Q}$ is ours, we can ask $\mathbf{Q}$ to be invertible. Then, multiplying (4) by $\mathbf{Q}^{-1}$ gives $\mathbf{Q}^{-1}\mathbf{Q}\tilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\tilde{\mathbf{x}}$ or

$$\tilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\tilde{\mathbf{x}}. \tag{5}$$

Given the $\mathbf{A}$ matrix, the idea is to try to find a $\mathbf{Q}$ matrix so that

$$\boxed{\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}} \tag{6}$$

is diagonal because then the differential equations within (5) will be *uncoupled.* If there does exist such a $\mathbf{Q}$ we say that $\mathbf{A}$ is **diagonalizable** and that $\mathbf{Q}$ **diagonalizes** $\mathbf{A}$.

Two questions present themselves: given $\mathbf{A}$, does there exist such a $\mathbf{Q}$ and, if so, how do we find it? (There is also a question of uniqueness, but we are not especially interested in whether or not $\mathbf{Q}$ is unique; we'll be happy to find *any* $\mathbf{Q}$ that diagonalizes $\mathbf{A}$.)

THEOREM 11.4.1 *Diagonalization*
Let $\mathbf{A}$ be $n \times n$.

1. $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors.
2. If $\mathbf{A}$ has n LI eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ and we make these the columns of $\mathbf{Q}$, so that $\mathbf{Q} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$, then $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal and the jth diagonal element of $\mathbf{D}$ is the jth eigenvalue of $\mathbf{A}$.

Proof: First, by $\mathbf{Q} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$ we mean that $\mathbf{Q}$ is partitioned into columns, the columns being $\mathbf{e}_1, \dots, \mathbf{e}_n$.

Let us prove that if $\mathbf{A}$ is diagonalizable, then it has n LI eigenvectors. If $\mathbf{A}$ is diagonalizable, then there is an invertible matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_n \end{bmatrix}. \tag{7}$$

Pre-multiplying both sides of (7) by $\mathbf{Q}$ gives $\mathbf{AQ} = \mathbf{QD}$:

$$\mathbf{AQ} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_n \end{bmatrix}$$

$$= \begin{bmatrix} d_1 q_{11} & \cdots & d_n q_{1n} \\ \vdots & & \vdots \\ d_1 q_{n1} & \cdots & d_n q_{nn} \end{bmatrix} \equiv [d_1 \mathbf{q}_1, \dots, d_n \mathbf{q}_n], \tag{8}$$

where the vector $\mathbf{q}_j$ simply denotes the jth column of $\mathbf{Q}$. Alternatively,

$$\mathbf{AQ} = \mathbf{A}[\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n] = [\mathbf{Aq}_1, \mathbf{Aq}_2, \dots, \mathbf{Aq}_n] \tag{9}$$

and, comparing (8) and (9), we see that

$$\mathbf{Aq}_1 = d_1 \mathbf{q}_1, \quad \dots, \quad \mathbf{Aq}_n = d_n \mathbf{q}_n. \tag{10}$$

Does (10) imply that the $\mathbf{q}_j$'s are eigenvectors of $\mathbf{A}$? Only if we can be sure that they are nonzero. Since we have assumed $\mathbf{A}$ to be diagonalizable, $\mathbf{Q}$ must be invertible. Hence, none of its columns $\mathbf{q}_j$ can be $\mathbf{0}$. Thus, the d_j's and $\mathbf{q}_j$'s are the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{A}$. Furthermore, the rank of $\mathbf{Q}$ must be n since Q is to be invertible, so (Theorem 10.5.2) its columns must be LI.

Thus far we have proved half of item 1, that if $\mathbf{A}$ is diagonalizable, then it has n LI eigenvectors. In doing so we have also proved item 2. It remains to prove the rest of 1, that if $\mathbf{A}$ has n LI eigenvectors, then it is diagonalizable. To do so, let us take $\mathbf{Q}$ to be made up of columns which are the eigenvectors of $\mathbf{A}$, so $\mathbf{Q} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$. Then

$$\mathbf{AQ} = [\mathbf{Ae}_1, \dots, \mathbf{Ae}_n] = [\lambda_1 \mathbf{e}_1, \dots, \lambda_n \mathbf{e}_n] = \begin{bmatrix} \lambda_1 e_{11} & \cdots & \lambda_n e_{n1} \\ \vdots & & \vdots \\ \lambda_1 e_{1n} & \cdots & \lambda_n e_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} e_{11} & \cdots & e_{n1} \\ \vdots & & \vdots \\ e_{1n} & \cdots & e_{nn} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_n \end{bmatrix} = \mathbf{QD}. \tag{11}$$

Finally, $\mathbf{Q}$ is invertible since its columns are LI, so pre-multiplying (11) by $\mathbf{Q}^{-1}$ gives $\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{D}$. Hence $\mathbf{A}$ is diagonalizable, and the proof is complete. ■

Since the columns of $\mathbf{Q}$ are LI eigenvectors of $\mathbf{A}$, $\mathbf{Q}$ is called a **modal matrix** of $\mathbf{A}$.

Theorem 11.4.1 relates the diagonalizability of $\mathbf{A}$ to the eigenvectors of $\mathbf{A}$. With the help of Theorem 11.4.2, we will be able to relate the diagonalizability of $\mathbf{A}$ to the eigenvalues of $\mathbf{A}$ as well. When that is done we will turn to applications.

THEOREM 11.4.2 *Distinct Eigenvalues, LI Eigenvectors*
If an $n \times n$ matrix $\mathbf{A}$ has distinct eigenvalues $\lambda_1, \dots, \lambda_n$, then the corresponding eigenvectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ are LI.

Proof: We need to show that

$$c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n = \mathbf{0} \tag{12}$$

holds only if $c_1 = c_2 = \cdots = c_n = 0$. Multiplying (12) by $\mathbf{A}$ and noting that $\mathbf{A}\mathbf{e}_j = \lambda_j\mathbf{e}_j$, we have

$$c_1\lambda_1\mathbf{e}_1 + c_2\lambda_2\mathbf{e}_2 + \cdots + c_n\lambda_n\mathbf{e}_n = \mathbf{0}. \tag{13}$$

Repeating the process gives

$$\begin{gathered} c_1\lambda_1^2\mathbf{e}_1 + c_2\lambda_2^2\mathbf{e}_2 + \cdots + c_n\lambda_n^2\mathbf{e}_n = \mathbf{0}, \\ \vdots \\ c_1\lambda_1^{n-1}\mathbf{e}_1 + c_2\lambda_2^{n-1}\mathbf{e}_2 + \cdots + c_n\lambda_n^{n-1}\mathbf{e}_n = \mathbf{0}. \end{gathered} \tag{14}$$

Expressed in matrix form,

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \begin{bmatrix} c_1\mathbf{e}_1 \\ c_2\mathbf{e}_2 \\ \vdots \\ c_n\mathbf{e}_n \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{15}$$

The determinant of the coefficient matrix is a *Vandermonde determinant*, which (see Exercise 17 in Section 10.4) is nonzero if the λ_j's are distinct. Since λ_j's are indeed distinct by assumption, (15) admits the unique trivial solution $c_1\mathbf{e}_1 = \mathbf{0}, c_2\mathbf{e}_2 = \mathbf{0}, \dots, c_n\mathbf{e}_n = \mathbf{0}$, and since the $\mathbf{e}_j$'s are nonzero (because they are eigenvectors) it follows that $c_1 = c_2 = \cdots = c_n = 0$, so $\mathbf{e}_1, \dots, \mathbf{e}_n$ are LI. ■

From Theorems 11.4.1 and 11.4.2 we can draw the following conclusion.

THEOREM 11.4.3 *Diagonalizability*
If an $n \times n$ matrix has n distinct eigenvalues, then it is diagonalizable.

As usual, be careful not to read converses into theorems when they are not stated. Specifically, Theorem 11.4.3 does *not* say that an $n \times n$ matrix is diagonalizable *if and only if* it has n distinct eigenvalues.

Consider an application.

EXAMPLE 1. *A Problem in Chemical Kinetics.* We consider, here, a special class of chemical reactions known as *first-order reactions*. These reactions are governed by systems of linear, coupled, first-order ordinary differential equations. Specifically, suppose that $X_1, \dots, X_n$ are the chemical names of n reacting species (elements or molecules), that $x_j(t)$ denotes the concentration of X_j (in suitable units) as a function of the time t, and that the *rate constant* for the conversion of X_i to X_j is the positive constant k_{ji}. For a two-component reaction, for example, denoted schematically in Fig. 1, this means that

$$x_1' = -k_{21}x_1 + k_{12}x_2, \tag{16a}$$

$$x_2' = k_{21}x_1 - k_{12}x_2. \tag{16b}$$

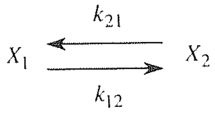


Figure 1. Two-component reaction.

The first term on the right-hand side of (16a) accounts for the loss of X_1 due to the $X_1 \to X_2$ reaction; it is proportional to the concentration of X_1, namely x_1, and the constant of proportionality is the relevant rate constant k_{21}. The second term on the right-hand side of (16a) accounts for the rate of gain of X_1 due to the reverse reaction $X_2 \to X_1$. A similar accounting holds for the terms in (16b).

The difficulty in solving (16), and similar systems for n-component reactions where $n > 2$, is due to the coupling. Equations (16) are coupled due to the $k_{12}x_2$ term in (16a) and the $k_{21}x_1$ term in (16b).

Let us solve (16) by diagonalization, if that is possible. In matrix form (16) is

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} -k_{21} & k_{12} \\ k_{21} & -k_{12} \end{bmatrix}. \tag{17}$$

The eigenvalues and eigenvectors of $\mathbf{A}$ are readily found to be

$$\lambda_1 = 0, \; \mathbf{e}_1 = \alpha \begin{bmatrix} k_{12} \\ k_{21} \end{bmatrix}; \qquad \lambda_2 = -(k_{12} + k_{21}), \; \mathbf{e}_2 = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{18}$$

The λ_j's are distinct, because $k_{12} > 0$ and $k_{21} > 0$, so Theorem 11.4.3 guarantees that $\mathbf{A}$ is diagonalizable. Alternatively, observe that the $\mathbf{e}_j$'s are necessarily LI because for them to be LD we would need $k_{21} = -k_{12}$, which is impossible since $k_{12} > 0$ and $k_{21} > 0$. Their linear independence implies that $\mathbf{A}$ is diagonalizable, by Theorem 11.4.1.

Thus, if we set $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where (with $\alpha = \beta = 1$, say)

$$\mathbf{Q} = [\mathbf{e}_1, \mathbf{e}_2] = \begin{bmatrix} k_{12} & 1 \\ k_{21} & -1 \end{bmatrix}, \tag{19}$$

then the preceding analysis assures us that

$$\tilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\tilde{\mathbf{x}} = \mathbf{D}\tilde{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tilde{\mathbf{x}}. \tag{20}$$

Thus, we have the *uncoupled* system (which, of course, is our objective)

$$\begin{aligned} \tilde{x}_1' &= \lambda_1 \tilde{x}_1, \\ \tilde{x}_2' &= \lambda_2 \tilde{x}_2, \end{aligned} \tag{21}$$

the general solution of which is

$$\begin{aligned} \tilde{x}_1 &= C_1 e^{\lambda_1 t} = C_1, \\ \tilde{x}_2 &= C_2 e^{\lambda_2 t} = C_2 e^{-(k_{12}+k_{21})t}. \end{aligned} \tag{22}$$

Finally, putting these expressions into $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ gives

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} k_{12} & 1 \\ k_{21} & -1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 e^{-(k_{12}+k_{21})t} \end{bmatrix} \tag{23}$$

or

$$\begin{aligned} x_1(t) &= C_1 k_{12} + C_2 e^{-(k_{12}+k_{21})t}, \\ x_2(t) &= C_1 k_{21} - C_2 e^{-(k_{12}+k_{21})t}. \end{aligned} \tag{24}$$

COMMENT 1. Here we have emphasized the mathematics rather than the chemistry and have assumed the rate constants to be known. A problem of importance to the chemist is the determination of those constants. Such determination normally involves a blend of the foregoing theory with suitable experiments.

COMMENT 2. The numbering of the eigenvalues and eigenvectors is immaterial. For instance, we could just as well take $\lambda_1 = -(k_{12} + k_{21})$ and $\lambda_2 = 0$. The final result, (24), would be the same. ∎

Theorem 11.4.3 revealed that diagonalizability is the typical case, the generic case, because an nth degree algebraic equation (namely, the characteristic equation of $\mathbf{A}$) typically has distinct roots. Furthermore, every *symmetric* matrix is diagonalizable:

THEOREM 11.4.4 *Symmetric Matrices*
Every symmetric matrix is diagonalizable.

Proof: Theorem 11.4.1 states that $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors, and Theorem 11.3.4 assures us that every $n \times n$ symmetric matrix has n orthogonal (and hence LI) eigenvectors. ∎

Suppose that for a symmetric matrix $\mathbf{A}$ we use the *normalized* eigenvectors of $\mathbf{A}$ to form its modal matrix $\mathbf{Q}$ so that

$$\mathbf{Q} = [\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n]. \tag{25}$$

Then, observe that

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \begin{bmatrix} \hat{\mathrm{e}}_1^{\mathrm{T}} \\ \vdots \\ \hat{\mathrm{e}}_n^{\mathrm{T}} \end{bmatrix} [\hat{\mathrm{e}}_1, \dots, \hat{\mathrm{e}}_n] = \begin{bmatrix} \hat{\mathrm{e}}_1^{\mathrm{T}}\hat{\mathrm{e}}_1 & \cdots & \hat{\mathrm{e}}_1^{\mathrm{T}}\hat{\mathrm{e}}_n \\ \vdots & & \vdots \\ \hat{\mathrm{e}}_n^{\mathrm{T}}\hat{\mathrm{e}}_1 & \cdots & \hat{\mathrm{e}}_n^{\mathrm{T}}\hat{\mathrm{e}}_n \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\mathrm{e}}_1 \cdot \hat{\mathrm{e}}_1 & \cdots & \hat{\mathrm{e}}_1 \cdot \hat{\mathrm{e}}_n \\ \vdots & & \vdots \\ \hat{\mathrm{e}}_n \cdot \hat{\mathrm{e}}_1 & \cdots & \hat{\mathrm{e}}_n \cdot \hat{\mathrm{e}}_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \mathbf{I} \tag{26}$$

so that

$$\boxed{\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}.} \tag{27}$$

Be sure to understand each step in (26). $\mathbf{Q}$ starts out as an $n \times n$ matrix, but when we partition it into columns, as $[\hat{\mathrm{e}}_1, \dots, \hat{\mathrm{e}}_n]$, it is then a $1 \times n$ matrix with elements $\hat{\mathrm{e}}_1, \dots, \hat{\mathrm{e}}_n$. To form $\mathbf{Q}^{\mathrm{T}}$, we make the jth column of $\mathbf{Q}$, namely $\hat{\mathrm{e}}_j$, the jth row of $\mathbf{Q}^{\mathrm{T}}$, and to put it into row format we need to write it as $\hat{\mathrm{e}}_j^{\mathrm{T}}$ rather than $\hat{\mathrm{e}}_j$. Thus, working with the partitioned $\mathbf{Q}$ and $\mathbf{Q}^{\mathrm{T}}$ matrices, the product to the right of the first equal sign in (26) is an $n \times 1$ matrix times a $1 \times n$ matrix, which product gives the $n \times n$ matrix to the right of the second equal sign.

Understand also that (27) has nothing to do with the $\hat{\mathrm{e}}_j$'s being eigenvectors. The steps in (26) rely only on the fact that the columns of $\mathbf{Q}$ are ON. *Any* square matrix, the columns of which are ON, satisfies (27) and is called an **orthogonal matrix.**[*] Of course, the property (27) is very nice because if we ever need the inverse of $\mathbf{Q}$ it is simply $\mathbf{Q}^{\mathrm{T}}$.

Getting back to diagonalization, note that if $\mathbf{A}$ is symmetric then, $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal whether or not the columns of the modal matrix $\mathbf{Q}$ are normalized. However, let us agree (at least within this text) to always normalize them, if $\mathbf{A}$ is symmetric, so as to have access to the property (27) if we need it.

Let us close with one more application.

EXAMPLE 2. *A Free-Vibration Problem.* Consider a mass m constrained by two mechanical springs, of stiffnesses k_1 and k_2, as sketched in Fig. 2. Imagine Fig. 2 as a view looking down on the apparatus, which lies in a horizontal plane on a frictionless table. In the configuration shown, the springs are neither stretched nor compressed, and m is at rest in static equilibrium. However, if some initial displacement and/or velocity is imparted to m, some motion, no doubt vibrational, will result, and it is that motion that we wish to determine.

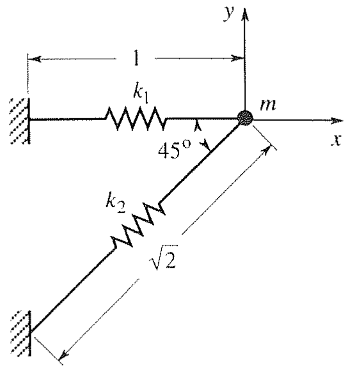


Figure 2. Mass-spring system; view from above.

The first step in the formulation is to introduce a coordinate system. A reasonable choice is the Cartesian system shown in Fig. 2, with its origin at the equilibrium position of the mass m (which we regard as a "point mass").

If m is at some point x, y other than the origin, then one or both springs will be stretched or compressed and will exert forces $\mathbf{F}_1$ and $\mathbf{F}_2$ on m (Fig.2). The magnitude

[*] Recall our encounter with orthogonal matrices in the optional Section 10.7.

of $\mathbf{F}_1$ is

$$\begin{aligned} \|\mathbf{F}_1\| &= k_1 \text{ times the stretch in spring \# 1} \\ &= k_1 \left\{ \sqrt{[x-(-1)]^2 + (y-0)^2} - 1 \right\} \\ &= k_1 \left\{ \sqrt{(x+1)^2 + y^2} - 1 \right\}. \end{aligned} \tag{28}$$

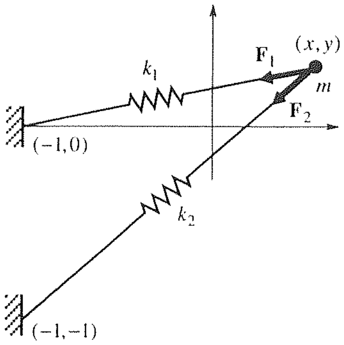


Figure 3. The forces on m.

If we multiply this magnitude by a unit vector directed from (x, y) toward $(-1, 0)$, we will have $\mathbf{F}_1$. The vector from (x, y) to $(-1, 0)$ is $(-1-x)\hat{\mathbf{i}} + (0-y)\hat{\mathbf{j}}$, where $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ are unit base vectors in the x, y directions, respectively. Normalizing that vector gives the desired unit vector

$$\hat{\mathbf{F}}_1 = -\frac{(1+x)\hat{\mathbf{i}} + y\hat{\mathbf{j}}}{\sqrt{(1+x)^2 + y^2}}, \tag{29}$$

so

$$\mathbf{F}_1 = \|\mathbf{F}_1\|\,\hat{\mathbf{F}}_1 = -k_1 \left[\frac{\sqrt{(x+1)^2 + y^2} - 1}{\sqrt{(x+1)^2 + y^2}} \right] \left[(x+1)\hat{\mathbf{i}} + y\hat{\mathbf{j}} \right]. \tag{30}$$

In like manner, we find that

$$\mathbf{F}_2 = \|\mathbf{F}_2\|\,\hat{\mathbf{F}}_2 = -k_2 \left[\frac{\sqrt{(x+1)^2 + (y+1)^2} - \sqrt{2}}{\sqrt{(x+1)^2 + (y+1)^2}} \right] \left[(x+1)\hat{\mathbf{i}} + (y+1)\hat{\mathbf{j}} \right]. \tag{31}$$

According to Newton's second law,

$$mx'' = F_x \quad \text{and} \quad my'' = F_y, \tag{32}$$

where F_x is the sum of the x components of $\mathbf{F}_1$ and $\mathbf{F}_2$, and F_y is the sum of the y components of $\mathbf{F}_1$ and $\mathbf{F}_2$. Thus, the governing equations of motion are

$$\begin{aligned} mx'' = &-k_1 \left[\frac{\sqrt{(x+1)^2 + y^2} - 1}{\sqrt{(x+1)^2 + y^2}} \right] (x+1) \\ &-k_2 \left[\frac{\sqrt{(x+1)^2 + (y+1)^2} - \sqrt{2}}{\sqrt{(x+1)^2 + (y+1)^2}} \right] (x+1), \end{aligned} \tag{33a}$$

$$\begin{aligned} my'' = &-k_1 \left[\frac{\sqrt{(x+1)^2 + y^2} - 1}{\sqrt{(x+1)^2 + y^2}} \right] y \\ &-k_2 \left[\frac{\sqrt{(x+1)^2 + (y+1)^2} - \sqrt{2}}{\sqrt{(x+1)^2 + (y+1)^2}} \right] (y+1). \end{aligned} \tag{33b}$$

The latter, coupled, nonlinear differential equations are, clearly, quite intractable analytically. Two possibilities present themselves. First, if we assign numerical values to $m, k_1, k_2, x(0), y(0), x'(0), y'(0)$, then we can generate $x(t)$ and $y(t)$ by one of the numerical methods studied in Chapter 6 (such as fourth-order Runge–Kutta integration) or using computer software (such as the *Maple* dsolve command).

Second, we can limit our attention to small motions, motions that remain close to the equilibrium position at the origin: $|x| \ll 1$ and $|y| \ll 1$. In that case we can simplify (33) in essentially the same way that we can simplify the nonlinear differential equation

$$x'' + \frac{g}{l}\sin x = 0 \tag{34}$$

governing the motion of a pendulum (Fig. 4): for small motions, near the equilibrium point $x = 0$,

$$\begin{aligned} \sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \qquad \text{(Taylor series)} \\ &\approx x \end{aligned}$$

for $|x| \ll 1$ so (34) can be approximated by the simple linear equation

$$x'' + \frac{g}{l}x = 0. \tag{35}$$

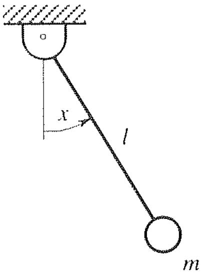


Figure 4. Pendulum, equations (34) and (35).

We will follow the same steps for (33), but instead of the Taylor series in one variable (about $x = 0$),

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots, \tag{36}$$

we will need the Taylor series in two variables (about $x = y = 0$):*

$$\begin{aligned} f(x,y) &= f(0,0) + f_x(0,0)x + f_y(0,0)y \\ &+ \frac{1}{2!}\left[f_{xx}(0,0)x^2 + 2f_{xy}(0,0)xy + f_{yy}(0,0)y^2\right] + \cdots \end{aligned} \tag{37}$$

because the right-hand sides of (33a,b) are functions of x and y. First, let $f(x,y)$ in (37) be the right-hand side of (33a). Then (37) gives

$$\begin{aligned} f(x,y) &= 0 + \left(-k_1 - \frac{k_2}{2}\right)x + \left(-\frac{k_2}{2}\right)y + \cdots \\ &\approx -\left(k_1 + \frac{k_2}{2}\right)x - \frac{k_2}{2}y. \end{aligned} \tag{38}$$

Next, let $f(x,y)$ in (37) be the right-hand side of (33b). Then (37) gives

$$\begin{aligned} f(x,y) &= 0 - \frac{k_2}{2}x - \frac{k_2}{2}y + \cdots \\ &\approx -\frac{k_2}{2}x - \frac{k_2}{2}y. \end{aligned} \tag{39}$$

With these approximations (linearizations) of the right-hand sides of (33a,b), we have the *linearized* equations

$$mx'' = -\left(k_1 + \frac{k_2}{2}\right)x - \frac{k_2}{2}y, \tag{40a}$$

$$my'' = -\frac{k_2}{2}x - \frac{k_2}{2}y \tag{40b}$$

*Taylor series in more than one variable is discussed in Chapter 13.

or

$$\mathbf{x}'' + \mathbf{A}\mathbf{x} = \mathbf{0}, \tag{41}$$

where

$$\mathbf{x} = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \dfrac{2k_1 + k_2}{2m} & \dfrac{k_2}{2m} \\ \dfrac{k_2}{2m} & \dfrac{k_2}{2m} \end{bmatrix}. \tag{42}$$

Let us solve (41) by diagonalization. Observe that $\mathbf{A}$ is symmetric (even though there is no "physical" symmetry to be seen in Fig. 2). For definiteness, let us set

$$m = 1, \; k_1 = 3, \; k_2 = 4. \tag{43}$$

Then the eigenvalues and normalized eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 1, \; \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ -2 \end{bmatrix}; \qquad \lambda_2 = 6, \; \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \tag{44}$$

With

$$\mathbf{Q} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2] = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}, \tag{45}$$

set $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$ in (41). Thus,

$$\mathbf{Q}\widetilde{\mathbf{x}}'' + \mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} = \mathbf{0} \tag{46}$$

and hence

$$\widetilde{\mathbf{x}}'' + \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} = \mathbf{0} \tag{47}$$

or

$$\widetilde{\mathbf{x}}'' + \mathbf{D}\widetilde{\mathbf{x}} = \mathbf{0}, \tag{48}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 6 \end{bmatrix}. \tag{49}$$

In scalar form, (48) gives the *uncoupled* equations

$$\begin{aligned} \widetilde{x}'' + \widetilde{x} &= 0, \\ \widetilde{y}'' + 6\widetilde{y} &= 0 \end{aligned} \tag{50}$$

with general solution [expressed in the $A \sin(\omega t + \phi)$ form]:

$$\begin{aligned} \widetilde{x} &= A_1 \sin(t + \phi_1), \\ \widetilde{y} &= A_2 \sin(\sqrt{6}\, t + \phi_2), \end{aligned} \tag{51}$$

where the amplitudes A_1, A_2 and phase angles ϕ_1, ϕ_2 are the four constants of integration. To return to the original x, y variables, write

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{Q}\widetilde{\mathbf{x}} &= \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \widetilde{x} \\ \widetilde{y} \end{bmatrix} \\ &= A_1 \begin{bmatrix} \dfrac{1}{\sqrt{5}} \sin(t + \phi_1) \\ -\dfrac{2}{\sqrt{5}} \sin(t + \phi_1) \end{bmatrix} + A_2 \begin{bmatrix} \dfrac{2}{\sqrt{5}} \sin(\sqrt{6}\, t + \phi_2) \\ \dfrac{1}{\sqrt{5}} \sin(\sqrt{6}\, t + \phi_2) \end{bmatrix} \\ &= A_1 \hat{\mathbf{e}}_1 \sin(t + \phi_1) + A_2 \hat{\mathbf{e}}_2 \sin(\sqrt{6}\, t + \phi_2) \end{aligned} \tag{52}$$

or, in scalar form,

$$\begin{aligned} x(t) &= C_1 \sin(t+\phi_1) + 2C_2 \sin(\sqrt{6}\,t+\phi_2), \\ y(t) &= -2C_1 \sin(t+\phi_1) + C_2 \sin(\sqrt{6}\,t+\phi_2), \end{aligned} \tag{53}$$

where C_1 is $A_1/\sqrt{5}$ and C_2 is $A_2/\sqrt{5}$, for brevity.

COMMENT 1. It is seen from (52) that the general solution is a linear combination of two orthogonal modes, as in Example 2 of Section 11.3. The low mode is a vibration along the $\hat{\mathbf{e}}_1$ direction, at a frequency that is the square root of λ_1, and the high mode is a vibration along the $\hat{\mathbf{e}}_2$ direction, at a frequency that is the square root of λ_2, as summarized in Fig. 5. In this example the orthogonality of the modes is geometric since the low- and high-mode motions are 90° apart; in Example 2 of Section 11.3, the orthogonality is more mathematical ($\mathbf{e}_1 \cdot \mathbf{e}_2$ being zero) than geometric.

COMMENT 2. Why are the directions of the low and high modes, shown in Fig. 5, physically reasonable? Recall that the natural frequency of the classical harmonic mechanical oscillator, governed by the equation $mx'' + kx = 0$, is $\sqrt{k/m}$, which increases as the stiffness k increases. It should be clear intuitively that the $\mathbf{e}_2$ axis is the line of maximum stiffness (of the two-spring system) and the $\mathbf{e}_1$ axis is the line of minimum stiffness, at least roughly speaking. Strikingly, the mathematics reveals that these directions are necessarily 90° apart.

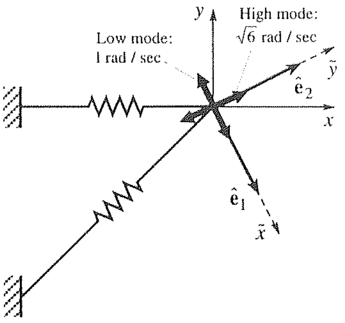


Figure 5. The orthogonal modes.

COMMENT 3. Why do we show the positive $\tilde{x}$ and $\tilde{y}$ coordinate axes as being in the $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ directions, respectively, in Fig. 5? Because if we set $\tilde{x} = 1$ and $\tilde{y} = 0$ in

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}$$

we get $[x, y]^{\mathrm{T}} = [1/\sqrt{5}, -2/\sqrt{5}]^{\mathrm{T}} = \hat{\mathbf{e}}_1$, and if we set $\tilde{x} = 0$ and $\tilde{y} = 1$ we get $[x, y]^{\mathrm{T}} = [2/\sqrt{5}, 1/\sqrt{5}]^{\mathrm{T}} = \hat{\mathbf{e}}_2$. In fact, if you studied the optional Section 10.7, you will appreciate that the effect of the change of variables $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where $\mathbf{Q}$ is an *orthogonal* matrix with its determinant equal to $+1$, as here, is a pure *rotation* of axes. Thus, we have the vivid visual image of the elements of the coupling matrix varying as we rotate the Cartesian coordinate system (somewhat like looking into a kaleidoscope), until the off-diagonal terms become zero and the equations uncouple.

COMMENT 4. In principle, it would have been best to choose the $\tilde{x}, \tilde{y}$ coordinate system in the first place, but its orientation was not known. Thus, we chose any x, y system, to get started, and then used the method of diagonalization to find the optimal $\tilde{x}, \tilde{y}$ coordinate system. ▮

Closure. From a mathematical viewpoint, this section is about finding an $n \times n$ matrix $\mathbf{Q}$, given an $n \times n$ matrix $\mathbf{A}$, such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal. We find that in the generic case $\mathbf{A}$ is diagonalizable: it is diagonalizable if and only if it has n LI eigenvectors, and it is diagonalizable if it has n distinct eigenvalues or is symmetric. $\mathbf{Q}$ can be made up of columns which are the eigenvectors of $\mathbf{A}$, and the diagonal elements of $\mathbf{D}$ are the corresponding eigenvalues. In the event that $\mathbf{A}$

is symmetric, we suggest always normalizing the eigenvectors that are the columns of $\mathbf{Q}$, so that $\mathbf{Q}$ will admit the useful property $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}$; that is, so that $\mathbf{Q}$ will be an orthogonal matrix.

From an applications standpoint, we look only at the use of digonalization in uncoupling systems of coupled differential equations, but additional applications are to be found in the exercises and in the next two sections.

It turns out that even if an $n \times n$ matrix $\mathbf{A}$ cannot be diagonalized, it can be triangularized. That is, a generalized modal matrix $\mathbf{P}$ can be found for $\mathbf{A}$ so that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} \tag{54}$$

is triangular. Called the **Jordan normal form**, or simply the **Jordan form**, for $\mathbf{A}$, $\mathbf{J}$ is upper triangular, with zeros above its main diagonal – except for 1's immediately above one or more diagonal elements. This case is discussed briefly in the exercises.

EXERCISES 11.4

1. Diagonalize each of the given $\mathbf{A}$ matrices. That is, determine matrices $\mathbf{Q}$ and $\mathbf{D}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal. Also, work out $\mathbf{Q}^{-1}$ and verify that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is diagonal and that its diagonal elements are the eigenvalues of $\mathbf{A}$. If $\mathbf{A}$ is not diagonalizable, state that and give the reason.

(a) $\begin{bmatrix} 2 & -3 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 4 \\ -1 & -2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(i) $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$

2. Use the method of diagonalization to obtain the general solution of the given system of differential equations, where primes denote d/dt.

(a) $x' = x + y$
$y' = x + y$

(b) $x' = x + 4y$
$y' = x + y$

(c) $x'' = 2x + 4y$
$y'' = x - y$

(d) $x' + 2x + y = 0$
$y' + x + 2y + z = 0$
$z' + y + 2z = 0$

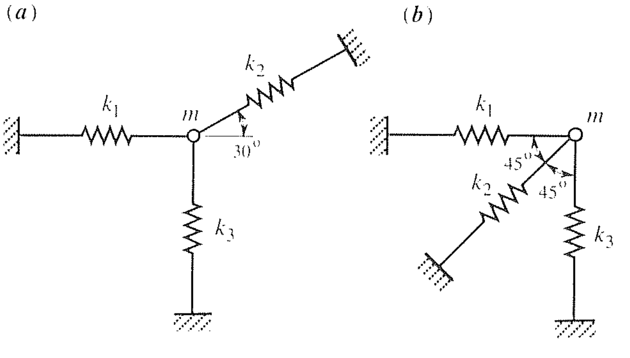
(e) $x' = 4x + y + 3z$
$y' = x - z$
$z' = 3x - y + 4z$

(f) $x'' = -y + z$
$y'' = -x - 3y - 4z$
$z'' = x - 4y - 3z$

3. Can a singular matrix be diagonalized? Explain.

4. We see from Fig. 5 that the line of action of the high mode falls between the two springs. Show that that situation holds for all possible combinations of stiffnesses k_1 and k_2.

5. Determine (as in Example 2) the natural frequencies and mode shapes for each of the systems shown below. You need carry only three or four significant figures. Each spring is of unit length.



$m = k_1 = k_2 = 1,\ k_3 = 10$ $m = k_1 = 4,\ k_2 = 3,\ k_3 = 1$

6. Show why the second equality in (9) is true.

7. (*Application to exponentiation*) Diagonalization can be helpful in raising a square matrix to a large power. Specifically, show that if $\mathbf{A}$ is diagonalizable so that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$, then

$$\mathbf{A}^m = \mathbf{Q}\mathbf{D}^m\mathbf{Q}^{-1}, \tag{7.1}$$

the advantage being that $\mathbf{D}^m$ is simply

$$\mathbf{D}^m = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}^m = \begin{bmatrix} \lambda_1^m & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^m \end{bmatrix}. \tag{7.2}$$

8. Use (7.1), above, to evaluate $\mathbf{A}^{1000}$, where

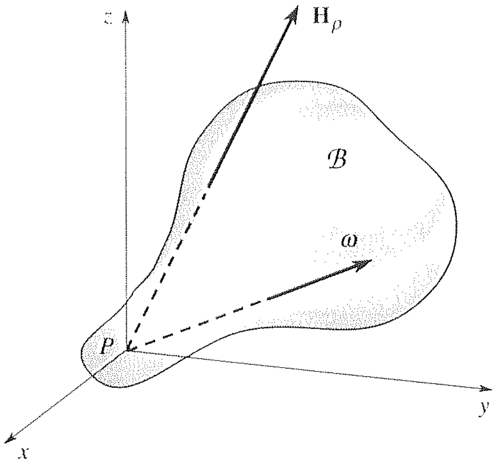
$$\text{(a) } \mathbf{A} = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix} \qquad \text{(b) } \mathbf{A} = \begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

9. (*Application to principal inertias and principal axes*) Two vectors of importance in studying the dynamics of a rigid body $\mathcal{B}$ are the *moment of momentum* $\mathbf{H}_p$ and the *angular velocity* ω, of $\mathcal{B}$. These are related according to $\mathbf{H}_p = \mathcal{I}\omega$, where $\mathcal{I}$ is the *inertia matrix*. Written out, we have

$$\begin{bmatrix} (H_p)_x \\ (H_p)_y \\ (H_p)_z \end{bmatrix} = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \tag{9.1}$$

where P is the origin of a Cartesian x, y, z coordinate system (see the figure), and

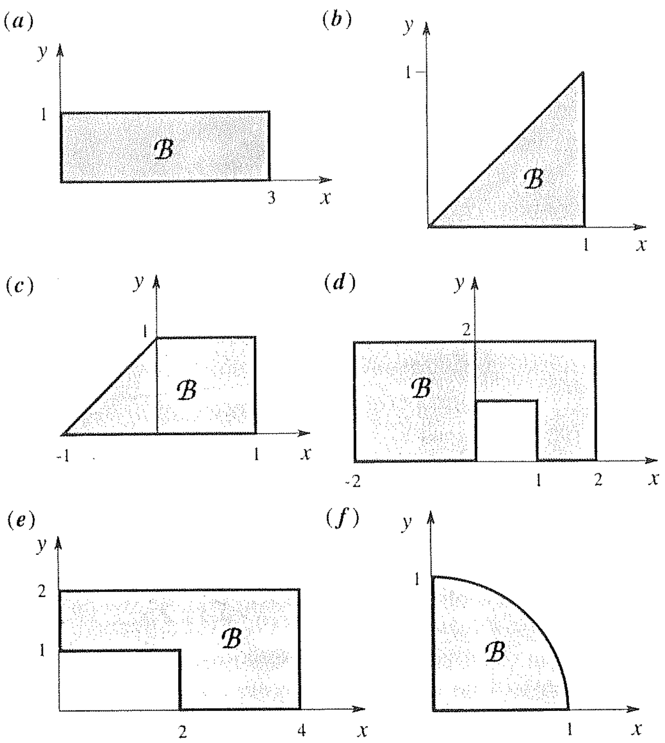
$$\begin{aligned} I_{xx} &= \int_{\mathcal{B}} (y^2 + z^2)\, dm, & I_{xy} = I_{yx} &= \int_{\mathcal{B}} xy\, dm \\ I_{yy} &= \int_{\mathcal{B}} (x^2 + z^2)\, dm, & I_{xz} = I_{zx} &= \int_{\mathcal{B}} xz\, dm \\ I_{zz} &= \int_{\mathcal{B}} (x^2 + y^2)\, dm, & I_{yz} = I_{zy} &= \int_{\mathcal{B}} yz\, dm. \end{aligned} \tag{9.2}$$



I_{xx}, I_{yy}, I_{zz} are known as the *moments of inertia* of $\mathcal{B}$ about the x, y, z axes, respectively, and I_{xy}, I_{xz}, I_{yz} are known as the *products of inertia* of $\mathcal{B}$; dm in (9.2) is "d(mass)." Now, the relation (9.1) and hence the subsequent dynamic analysis (which will be of no concern here) will be simplest if $\mathcal{I}$ is diagonal, i.e., if all of the products of inertia are zero. In general, it is too difficult to see, by inspection, how to orient the coordinate axes to achieve this result. Instead, we go ahead and choose *some* x, y, z reference frame, compute the nine inertia components, and then rotate to a new Cartesian $\tilde{x}, \tilde{y}, \tilde{z}$ frame so as to diagonalize $\mathcal{I}$. That is, if $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where $\mathbf{x} = [x, y, z]^{\mathrm{T}}$ and $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}, \tilde{z}]^{\mathrm{T}}$, then $\mathbf{H}_p = \mathbf{Q}\tilde{\mathbf{H}}_p$ and $\omega = \mathbf{Q}\tilde{\omega}$ so that $\mathbf{H}_p = \mathcal{I}\omega$ becomes

$$\mathbf{Q}\tilde{\mathbf{H}}_p = \mathcal{I}\mathbf{Q}\tilde{\omega} \quad \text{and} \quad \tilde{\mathbf{H}}_p = \left(\mathbf{Q}^{-1}\mathcal{I}\mathbf{Q}\right)\tilde{\omega}, \tag{9.3}$$

where $\mathbf{Q}^{-1}\mathcal{I}\mathbf{Q} = \tilde{\mathcal{I}}$ is diagonal. That such diagonalization is possible follows from the fact that $\mathcal{I}$ is *symmetric* since $I_{xy} = I_{yx}$, $I_{xz} = I_{zx}$, and $I_{yz} = I_{zy}$. $I_{\tilde{x}\tilde{x}}$, $I_{\tilde{y}\tilde{y}}$, and $I_{\tilde{z}\tilde{z}}$ are called the **principal inertias** of $\mathcal{B}$ (with respect to coordinates with origin at P), and the $\tilde{x}, \tilde{y}, \tilde{z}$ axes are called the **principal axes**. We now state the problem: compute the principal inertias and determine the principal axes for each of the following bodies; sketch the principal axes. In each case $\mathcal{B}$ can be assumed, for simplicity, to be infinitely thin, with mass density σ mass units per unit area.



10. (*Jordan form*) Recall that an $n \times n$ matrix $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors. Thus, not all matrices can be diagonalized. However, our experience in Chapters 8–11 has shown that triangular matrices are "almost as nice" as diagonal ones, and it turns out that even if $\mathbf{A}$ cannot be diagonalized, it can be triangularized.

More specifically, if $\mathbf{A}$ is diagonalizable and its modal matrix is $\mathbf{Q} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$, where $\mathbf{e}_j$'s are the LI eigenvectors of $\mathbf{A}$, then $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal, with d_{jj} equal to the jth eigenvalue of $\mathbf{A}$. But suppose $\mathbf{A}$ is *not* diagonalizable. Though it does not have n LI eigenvectors, it has n LI **generalized eigenvectors** $\mathbf{e}_1, \dots, \mathbf{e}_n$ (defined below) and if its **generalized modal matrix** is $\mathbf{P} = [\mathbf{e}_1, \dots, \mathbf{e}_n]$, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$, where $\mathbf{J}$ is the **Jordan form** for $\mathbf{A}$; $\mathbf{J}$ will be upper triangular, with the eigenvalues of $\mathbf{A}$ on its diagonal and zeros everywhere else except perhaps immediately above repeated eigenvalues. For details, we refer you to M. Greenberg, *Advanced Engineering Mathematics*, 1st ed. (Englewood Cliffs, NJ: Prentice Hall, 1988). Here, we will merely go through an example with you and, at the end, ask you to supply various steps.

Consider, as a representative case,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 2 & 0 \\ 0 & 3 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -3 & 5 \end{bmatrix}. \tag{10.1}$$

Calculation reveals these eigenvalues and eigenvectors:

$$\begin{aligned} \lambda_1 = \lambda_2 = \lambda_3 = 2, \qquad & \mathbf{e} = [1,0,0,0]^{\mathrm{T}} \equiv \mathbf{e}_1, \\ \lambda_4 = 5, \qquad & \mathbf{e} = [0,0,0,1]^{\mathrm{T}} \equiv \mathbf{e}_4. \end{aligned} \tag{10.2}$$

In this case the eigenvalue of multiplicity three, $\lambda = 2$, contributes only one LI eigenvector, instead of three, so we end up with only two LI eigenvectors, $\mathbf{e}_1$ and $\mathbf{e}_4$, instead of the four that are needed for diagonalization. Thus, $\mathbf{A}$ is not diagonalizable. But we can find "generalized eigenvectors" $\mathbf{e}_2$ and $\mathbf{e}_3$, associated with $\lambda = 2$, such that $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ are LI. Noting that $\mathbf{e}_1$ satisfies

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_1 = \mathbf{0}, \tag{10.3}$$

we introduce $\mathbf{e}_2, \mathbf{e}_3$ so as to satisfy

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_2 = \mathbf{e}_1, \tag{10.4}$$

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_3 = \mathbf{e}_2. \tag{10.5}$$

With λ_1 and $\mathbf{e}_1$ given in (10.2), solution of (10.4) and (10.5) by Gauss elimination gives

$$\mathbf{e}_2 = [\alpha, 1, 1, 2/3]^{\mathrm{T}}, \tag{10.6}$$

$$\mathbf{e}_3 = [\beta, \alpha + 2, \alpha + 1, 2\alpha/3 + 5/9]^{\mathrm{T}}, \tag{10.7}$$

where α, β are arbitrary. Take $\alpha = \beta = 0$, say. Then, with the generalized modal matrix

$$\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 2/3 & 5/9 & 1 \end{bmatrix}, \tag{10.8}$$

we obtain

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} = \mathbf{J}. \tag{10.9}$$

(a) Derive the eigenvalues and eigenvectors given in (10.2).
(b) Use Gauss elimination to derive (10.6) from (10.4), and (10.7) from (10.5).

(c) Use (10.3), (10.4), (10.5), and $(\mathbf{A} - \lambda_4 \mathbf{I})\mathbf{e}_4 = \mathbf{0}$ to show that $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ are LI, as claimed above (10.3). HINT: Suppose that

$$\alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3 + \alpha_4 \mathbf{e}_4 = \mathbf{0}. \tag{10.10}$$

Then multiply each term in (10.10) by $\mathbf{A} - \lambda_1 \mathbf{I}$. Then multiply each term in the resulting equation by $\mathbf{A} - \lambda_1 \mathbf{I}$. Again multiply each term in the resulting equation by $\mathbf{A} - \lambda_1 \mathbf{I}$. Explain why the resulting set of four equations implies that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$. NOTE: More generally, let λ be an eigenvalue of the given matrix $\mathbf{A}$, of multiplicity K, and let the eigenspace corresponding to λ be of dimension k, where $k < K$. Within that eigenspace there can be found k LI eigenvectors of $\mathbf{A}$, say $\mathbf{e}_1$ through $\mathbf{e}_k$. Vectors $\mathbf{e}_{k+1}$ through $\mathbf{e}_K$ satisfying

$$\begin{aligned}(\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_{k+1} &= \mathbf{e}_k, \\ (\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_{k+2} &= \mathbf{e}_{k+1}, \\ &\vdots \\ (\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_K &= \mathbf{e}_{K-1}\end{aligned} \tag{10.11}$$

are the **generalized eigenvectors** corresponding to λ.

(d) Consider the system of differential equations

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \tag{10.12}$$

where the prime denotes d/dt and $\mathbf{A}$ is given by (10.1). Under the change of variables $\mathbf{x} = \mathbf{P}\widetilde{\mathbf{x}}$, reduce (10.12) to the Jordan form

$$\widetilde{\mathbf{x}}' = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\widetilde{\mathbf{x}} = \mathbf{J}\widetilde{\mathbf{x}}, \tag{10.13}$$

where $\mathbf{J}$ is given in (10.9). Solve the triangular system (10.13) for $\widetilde{\mathbf{x}}(t)$, and thus obtain the general solution $\mathbf{x}(t) = \mathbf{P}\widetilde{\mathbf{x}}(t)$ of (10.12).

11.5 Application to First-Order Systems with Constant Coefficients (Optional)

In Section 11.4 we studied diagonalization, and showed how to use that method to uncouple and solve systems of differential equations. In this section we continue that discussion, but this time our emphasis is on the theory of differential equations rather than on diagonalization.

Consider the initial-value problem

$$\begin{aligned} x_1' &= a_{11}x_1 + \cdots + a_{1n}x_n + f_1(t); \qquad x_1(0) = c_1 \\ &\vdots \\ x_n' &= a_{n1}x_1 + \cdots + a_{nn}x_n + f_n(t); \qquad x_n(0) = c_n \end{aligned} \tag{1}$$

where the a_{ij}'s are constants. Or, in matrix form,

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{f}(t); \qquad \mathbf{x}(0) = \mathbf{c}. \tag{2}$$

Unlike Section 11.4, here we allow for forcing functions $f_1(t), \ldots, f_n(t)$.

To solve the first-order system (2) by diagonalization, we suppose that $\mathbf{A}$ has n LI eigenvectors, a modal matrix $\mathbf{Q}$, and eigenvalues $\lambda_1, \ldots, \lambda_n$ that are not necessarily distinct. Setting

$$\mathbf{x}(t) = \mathbf{Q}\widetilde{\mathbf{x}}(t), \tag{3}$$

(2) becomes

$$\mathbf{Q}\widetilde{\mathbf{x}}' = \mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} + \mathbf{f}(t); \qquad \mathbf{Q}\widetilde{\mathbf{x}}(0) = \mathbf{c} \tag{4}$$

or, equivalently,

$$\widetilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} + \mathbf{Q}^{-1}\mathbf{f}(t); \qquad \widetilde{\mathbf{x}}(0) = \mathbf{Q}^{-1}\mathbf{c}, \tag{5}$$

where $\mathbf{Q}^{-1}$ necessarily exists because the columns of $\mathbf{Q}$ are LI. We know that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \lambda_n \end{bmatrix}, \tag{6}$$

and if we denote $\mathbf{Q}^{-1}\mathbf{f}(t)$ as $\widetilde{\mathbf{f}}(t)$ and $\mathbf{Q}^{-1}\mathbf{c}$ as $\widetilde{\mathbf{c}}$, for brevity, then

$$\widetilde{\mathbf{x}}' = \mathbf{D}\widetilde{\mathbf{x}} + \widetilde{\mathbf{f}}(t); \qquad \widetilde{\mathbf{x}}(0) = \widetilde{\mathbf{c}} \tag{7}$$

or, returning to scalar form,

$$\begin{aligned} \widetilde{x}_1' &= \lambda_1 \widetilde{x}_1 + \widetilde{f}_1(t); \qquad & \widetilde{x}_1(0) &= \widetilde{c}_1 \\ &\vdots \\ \widetilde{x}_n' &= \lambda_n \widetilde{x}_n + \widetilde{f}_n(t); \qquad & \widetilde{x}_n(0) &= \widetilde{c}_n. \end{aligned} \tag{8}$$

Each of these equations is first-order linear, like equation (2) in Section 2.2, and its solution is given by (24) in Section 2.2.2 [with $x \to t, y \to \widetilde{x}_j, p(x) \to -\lambda_j$, and so on]:

$$\begin{aligned} \widetilde{x}_1(t) &= \widetilde{c}_1 e^{\lambda_1 t} + \int_0^t e^{\lambda_1 (t-\tau)} \widetilde{f}_1(\tau) d\tau, \\ &\vdots \\ \widetilde{x}_n(t) &= \widetilde{c}_n e^{\lambda_n t} + \int_0^t e^{\lambda_n (t-\tau)} \widetilde{f}_n(\tau) d\tau. \end{aligned} \tag{9}$$

If we define

$$e^{\mathbf{D}t} \equiv \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & e^{\lambda_n t} \end{bmatrix}, \tag{10}$$

then we can express (9) as

$$\widetilde{\mathbf{x}}(t) = e^{\mathbf{D}t}\widetilde{\mathbf{c}} + \int_0^t e^{\mathbf{D}(t-\tau)} \widetilde{\mathbf{f}}(\tau) d\tau. \tag{11}$$

Finally, since $\widetilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x}$, $\widetilde{\mathbf{c}} = \mathbf{Q}^{-1}\mathbf{c}$, and $\widetilde{\mathbf{f}} = \mathbf{Q}^{-1}\mathbf{f}$,

$$\boxed{\mathbf{x}(t) = \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}\mathbf{c} + \int_0^t \mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1}\mathbf{f}(\tau)d\tau} \tag{12}$$

is the unique solution of (1). Note the order: $e^{\mathbf{D}t}\widetilde{\mathbf{c}}$, not $\widetilde{\mathbf{c}}e^{\mathbf{D}t}$, in (11).

EXAMPLE 1. Use (12) to solve the system

$$\begin{aligned} x' &= x + 4y - 4t^2 - 3; & x(0) &= 2 \\ y' &= x + y - t^2 + 2t - 3; & y(0) &= 3. \end{aligned} \tag{13}$$

Then

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{f}(t) = \begin{bmatrix} -4t^2 - 3 \\ -t^2 + 2t - 3 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}. \tag{14}$$

The eigenvalues and eigenvectors of $\mathbf{A}$ are found to be

$$\lambda_1 = 3, \ \mathbf{e}_1 = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -1, \ \mathbf{e}_1 = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{15}$$

Since these $\mathbf{e}_j$'s are LI, $\mathbf{A}$ is diagonalizable so (12) applies. With

$$\mathbf{D} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{Q}^{-1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}, \tag{16}$$

(12) gives

$$\begin{aligned} \mathbf{x}(t) &= \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ &\quad + \int_0^t \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{3(t-\tau)} & 0 \\ 0 & e^{-(t-\tau)} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -4\tau^2 - 3 \\ -\tau^2 + 2\tau - 3 \end{bmatrix} d\tau \\ &= \begin{bmatrix} 4e^{3t} - 2e^{-t} \\ 2e^{3t} + e^{-t} \end{bmatrix} + \begin{bmatrix} 3 - \frac{3}{2}e^{3t} - \frac{3}{2}e^{-t} \\ t^2 - \frac{3}{4}e^{3t} + \frac{3}{4}e^{-t} \end{bmatrix} \end{aligned} \tag{17}$$

so

$$\begin{aligned} x(t) &= 3 + \frac{5}{2}e^{3t} - \frac{7}{2}e^{-t}, \\ y(t) &= t^2 + \frac{5}{4}e^{3t} + \frac{7}{4}e^{-t} \end{aligned} \tag{18}$$

is the desired solution. ▮

EXAMPLE 2. Can (12) be used for the system

$$x'' + 3x' + 2x - y = e^t \tag{19a}$$

$$y'' - y' + 5x' - x - 6y = t \text{ ?} \tag{19b}$$

Yes, if we can reduce (19) to a *first*-order system. How to do that is described in Section 3.9 (see Example 5, therein, and the paragraph preceding that example). Specifically, set $x' = u$ and $y' = v$. Then (19) can be re-expressed as

$$\begin{aligned} x' &= u, && \text{(by definition)} \\ u' &= -3u - 2x + y + e^t, && \text{[from (19a)]} \\ y' &= v, && \text{(by definition)} \\ v' &= v - 5u + x + 6y + t && \text{[from (19b)]} \end{aligned} \tag{20}$$

or, in matrix form, as

$$\begin{bmatrix} x' \\ u' \\ y' \\ v' \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -2 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & -5 & 6 & 1 \end{bmatrix} \begin{bmatrix} x \\ u \\ y \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ e^t \\ 0 \\ t \end{bmatrix}. \tag{21}$$

The latter is a first-order system, to which (12) could be applied. ▮

In deriving (12) an interesting quantity arose, the **exponential matrix function** $e^{\mathbf{B}}$, where $\mathbf{B} = \{b_{ij}\}$ is an $n \times n$ diagonal matrix:

$$\boxed{e^{\mathbf{B}} \equiv \begin{bmatrix} e^{b_{11}} & 0 & \cdots & 0 \\ 0 & e^{b_{22}} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & e^{b_{nn}} \end{bmatrix}.} \quad \text{(B diagonal)} \tag{22}$$

More generally, let $\mathbf{B}$ be any $n \times n$ matrix, not necessarily diagonal. Following the familiar Taylor series formula,

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \cdots,$$

it seems reasonable to define

$$\boxed{e^{\mathbf{B}} \equiv \mathbf{I} + \frac{1}{1!}\mathbf{B} + \frac{1}{2!}\mathbf{B}^2 + \cdots.} \tag{23}$$

Fine, but observe that the right-hand side of (23) is an *infinite series of matrices*, which we have not yet defined. Mimicing the usual definition of convergence for series of scalars, we define an infinite series of matrices $\sum_{j=1}^{\infty} \mathbf{A}_j$ as the limit of the sequence of partial sums $\mathbf{S}_N$, where $\mathbf{S}_N = \sum_{j=1}^{N} \mathbf{A}_j$. That is,

$$\sum_{j=1}^{\infty} \mathbf{A}_j \equiv \lim_{N\to\infty} \mathbf{S}_N. \tag{24}$$

The infinite series is said to **converge** if the limit on the right exists and to diverge if that limit does not exist. Finally, observe that $\lim_{N\to\infty} \mathbf{S}_N$ is the *limit of a sequence of matrices*, which we have not yet defined so we are not done.

Let $\mathbf{C}_1, \mathbf{C}_2, \ldots$ be a sequence of $m \times n$ matrices, with $(c_{ij})_n$ as the i, j element of $\mathbf{C}_n$. We say that the sequence **converges** to a matrix $\mathbf{C} = \{c_{ij}\}$ if

$$\lim_{n\to\infty} (c_{ij})_n = c_{ij}$$

for each i, j, and we denote such convergence by writing either $\lim_{n\to\infty} \mathbf{C}_n = \mathbf{C}$ or $\mathbf{C}_n \to \mathbf{C}$ as $n \to \infty$. If the sequence does not converge, then it is said to **diverge**.

EXAMPLE 3.

$$\begin{bmatrix} 2+\dfrac{3}{n} & e^{-n} \\ 1 & \dfrac{4}{n+2} \end{bmatrix} \to \begin{bmatrix} 2 & 0 \\ 1 & 0 \end{bmatrix}, \tag{25}$$

but

$$\begin{bmatrix} 7-3n & 2+e^{-4n} \\ 3 & \dfrac{1}{n} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 5 \\ \sin n & \dfrac{1}{n+2} \end{bmatrix} \tag{26}$$

diverge because $\lim_{n\to\infty}(7-3n)$ and $\lim_{n\to\infty} \sin n$ do not exist. ▮

With (23), let us return to (12) and consider the combination $\mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}$:

$$\begin{aligned} \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1} &= \mathbf{Q}\left(\mathbf{I} + \mathbf{D}t + \frac{1}{2!}\mathbf{D}^2t^2 + \cdots\right)\mathbf{Q}^{-1} \\ &= \mathbf{I} + \mathbf{QDQ}^{-1}t + \frac{1}{2!}\mathbf{QD}^2\mathbf{Q}^{-1}t^2 + \cdots \\ &= \mathbf{I} + \mathbf{QDQ}^{-1}t + \frac{1}{2!}\mathbf{QDQ}^{-1}\mathbf{QDQ}^{-1}t^2 + \cdots, \end{aligned} \tag{27}$$

and recalling that $\mathbf{QDQ}^{-1} = \mathbf{A}$, this series gives

$$\mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1} = \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots = e^{\mathbf{A}t}. \tag{28}$$

Similarly, $\mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1} = e^{\mathbf{A}(t-\tau)}$ so (12) can be expressed in the equivalent form

$$\boxed{\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{c} + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{f}(\tau)\,d\tau.} \tag{29}$$

How are we to evaluate the exponentials in (29)? We could use the series formula (23), but it is more efficient to reverse (28) and to use $e^{\mathbf{A}t} = \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}$ and $e^{\mathbf{A}(t-\tau)} = \mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1}$ because $e^{\mathbf{D}t}$ is given simply by (10), and similarly for $e^{\mathbf{D}(t-\tau)}$. That is, computationally, we continue to rely on (12) rather than (29). Nonetheless, (29) is of considerable interest because it shows how the solution

$$x(t) = e^{At}c + \int_0^t e^{A(t-\tau)} f(\tau)d\tau \tag{30}$$

of the single initial-value problem

$$x'(t) = Ax + f(t); \qquad x(0) = c \tag{31}$$

can be generalized to the solution (29) of (2) using the exponential matrix function.*

Closure. There were two objectives in this section. One was to derive the matrix solution (12) of the initial-value problem (2), and the other was to introduce the exponential matrix function $e^{\mathbf{B}}$ for any $n \times n$ matrix $\mathbf{B}$. In fact, one can introduce other functions of square matrices as well: sines, cosines, fractional powers, and so on. For discussion of such functions we refer you to Wylie and Barrett.†

Computer software. Using *Maple*, the exponential matrix function can be evaluated by means of the **exponential** command within the linalg package. For instance, to evaluate $e^{\mathbf{A}t}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix},$$

enter

```
with(linalg):
```

and return, then

$$\text{B} := \text{matrix}\ (2, 2,\ [t, 4 * t, t, t]):$$

and return. Then,

```
exponential(B);
```

and return gives the output

$$\begin{bmatrix} \frac{1}{2}e^{(-t)} + \frac{1}{2}e^{(3t)} & e^{(3t)} - e^{(-t)} \\ \frac{1}{4}e^{(3t)} - \frac{1}{4}e^{(-t)} & \frac{1}{2}e^{(-t)} + \frac{1}{2}e^{(3t)} \end{bmatrix}$$

Actually, when you enter the $\mathbf{B}$ matrix you may prefer to end with a semicolon rather than a colon because then the $\mathbf{B}$ matrix will be printed, and you can inspect it to see if its elements were entered correctly.

*Of course, it is more natural to write the first term on the right-hand side of (30) as ce^{At}, but we have written it as $e^{At}c$ to emphasize the correspondence between (29) and (30).

†C. R. Wylie and L. C. Barrett, *Advanced Engineering Mathematics*, 5th ed. (New York: McGraw–Hill, 1982).

EXERCISES 11.5

1. Solve, using (12). If (12) does not apply, explain why it does not. NOTE: You may need to first re-express the problem in the standard form (1).

(a) $x' = x + y$; $x(0) = 0$
$y' = x + y + e^{3t}$; $y(0) = 0$

(b) $x' = 2x + 4y + 1$; $x(0) = 0$
$y' = x - y$; $y(0) = 0$

(c) $x' = 2x - y - 3t + 1$; $x(0) = 1$
$y' = -x + 2y + 3$; $y(0) = 0$

(d) $x' = x + 2y - t - 1$; $x(0) = 0$
$y' = 4x + 8y - 4t - 8$; $y(0) = 3$

(e) $x' = \sin t - y$; $x(0) = 0$
$y' = -9x + 4$; $y(0) = 1$

(f) $x' = x - 8y$; $x(0) = 0$
$y' = -x - y - 3t^2$; $y(0) = 0$

(g) $x' - 3x + y = 4e^t$; $x(0) = 1$
$y' + 3x - y = 6$; $y(0) = -1$

(h) $x' - x - y = 3t$; $x(0) = 5$
$x' + y' - 5x - 2y = 5$; $y(0) = 3$

(i) $x'' + 3x' + 2x = e^t$; $x(0) = x'(0) = 0$

(j) $x'' + 2x' + x = 0$; $x(0) = 2$, $x'(0) = 9$

(k) $x_1' = 4x_1 + x_2 + 3x_3$; $x_1(0) = 1$
$x_2' = x_1 - x_3 + 6$; $x_2(0) = 0$
$x_3' = 3x_1 - x_2 + 4x_3$; $x_3(0) = 0$

(l) $x' = x + y + z - 1$; $x(0) = 0$
$y' = x + y + z$; $y(0) = 0$
$z' = x + y + z$; $z(0) = 0$

2. (a)–(l) Same as Exercise 1, but drop the initial conditions and find the general solution instead.

3. Evaluate $e^{\mathbf{A}}$ for the given $\mathbf{A}$ matrix. HINT: If $\mathbf{A}$ is diagonalizable you can use the formula

$$e^{\mathbf{A}} = \mathbf{Q}e^{\mathbf{D}}\mathbf{Q}^{-1} \tag{3.1}$$

[i.e., equation (28) with $t = 1$]. Whether or not $\mathbf{A}$ is diagonalizable, you can use the series definition

$$e^{\mathbf{A}} = \mathbf{I} + \frac{1}{1!}\mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \cdots. \tag{3.2}$$

Generally, (3.2) is unwieldy because one needs to sum an infinite series of matrices. However, if $\mathbf{A}$ happens to be nilpotent then the series reduces to a finite number of terms.

(a) $\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 4 \\ 9 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

(e) $\begin{bmatrix} 2 & 0 \\ 7 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (h) $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

(i) $\begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{bmatrix}$

(k) $\begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 1 & 0 \end{bmatrix}$

4. (a)–(l) Evaluate $e^{\mathbf{A}}$ using computer software, where $\mathbf{A}$ is given in the corresponding part of Exercise 3.

11.6 Quadratic Forms (Optional)

A function of the form

$$f(x_1, x_2) = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 \tag{1}$$

is called a **quadratic form** in the variables x_1 and x_2. Similarly,

$$f(x_1, x_2, x_3) = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + 2a_{23}x_2x_3 \tag{2}$$

is a quadratic form in x_1, x_2, x_3, and so on for any number of variables $x_1, \ldots, x_n$. The a_{ij}'s are constants (where a_{ij} is the coefficient of the product x_ix_j), and the 2's are included for our subsequent convenience.

It turns out that a quadratic form $f(x_1, \ldots, x_n)$ can be expressed concisely, in matrix notation, as

$$\boxed{f(x_1, \ldots, x_n) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x},} \tag{3}$$

where $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, and $\mathbf{A} = \{a_{ij}\}$ is an $n \times n$ matrix. For instance, suppose that $n = 2$ and

$$f(x_1, x_2) = 2x_1^2 - x_2^2 + 6x_1x_2. \tag{4}$$

Writing out the right-hand side of (3) gives

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} &= [x_1, x_2]\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [x_1, x_2]\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + a_{22}x_2^2 + (a_{12} + a_{21})x_1x_2. \end{aligned} \tag{5}$$

Comparing (4) and (5), we see that $a_{11} = 2, a_{22} = -1$, and $a_{12} + a_{21} = 6$. Since the sum $a_{12} + a_{21}$ is prescribed, but not a_{12} and a_{21} individually, we are free to ask a_{12} and a_{21} to be equal, so that $a_{12} = a_{21} = 3$. Then the right-hand side of (5) becomes $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$, as written in (1). The benefit in asking a_{12} to equal a_{21} is that then the $\mathbf{A}$ matrix in (3) is *symmetric*, and symmetric matrices have advantages over nonsymmetric matrices, as we have seen in Sections 11.3 and 11.4. Similarly for $n = 3, 4, \ldots$.

Thus, if we ask the $\mathbf{A}$ matrix in (3) to be symmetric, for convenience, then (3) gives (1) for $n = 2$, (2) for $n = 3$, and so on.

EXAMPLE 1. Let

$$f(x_1, \ldots, x_4) = x_2^2 + 5x_3^2 + 6x_1x_3 - x_2x_3 + 10x_3x_4.$$

Then $a_{22} = 1, a_{33} = 5, 2a_{13} = 6$ so $a_{13} = 3, 2a_{23} = -1$ so $a_{23} = -\frac{1}{2}, 2a_{34} = 10$ so $a_{34} = 5$, and all the other a_{ij}'s are zero. Thus

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 3 & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 3 & -\frac{1}{2} & 5 & 5 \\ 0 & 0 & 5 & 0 \end{bmatrix}. \quad \blacksquare$$

A quadratic form is said to be **canonical** if all mixed terms (such as x_1x_2, x_1x_3, and x_2x_3) are absent, that is, if $a_{ij} = 0$ for $i \neq j$. Thus,

$$f(x_1, \ldots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 \tag{6}$$

is canonical, and its associated matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & a_{nn} \end{bmatrix}$$

is diagonal. For instance, (4) is not canonical, but $f(x_1, x_2) = 5x_1^2 - 7x_2^2$ and $f(x_1, x_2, x_3) = 6x_2^2 + x_3^2$ are.

There is interest in being able to reduce a given quadratic form to canonical form (i.e., to its simplest form) through a linear change of variables

$$\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}} \tag{7}$$

from $x_1, \ldots, x_n$ to $\tilde{x}_1, \ldots, \tilde{x}_n$. Putting (7) into (3) gives

$$f = (\mathbf{Q}\tilde{\mathbf{x}})^{\mathrm{T}} \mathbf{A} (\mathbf{Q}\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^{\mathrm{T}} \left(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}\right) \tilde{\mathbf{x}}. \tag{8}$$

Thus, given $\mathbf{A}$, we wish to determine a $\mathbf{Q}$ matrix such that $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$ is diagonal, for then (8) will be canonical in the new variables $\tilde{x}_1, \ldots, \tilde{x}_n$. Theorem 11.4.1 tells us that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ will be a diagonal matrix, with its diagonal elements being the eigenvalues of $\mathbf{A}$, if $\mathbf{A}$ has n LI eigenvectors and if these eigenvectors are used as the columns of $\mathbf{Q}$.* However, (8) contains $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$, not $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$. But since $\mathbf{A}$ is symmetric it has n orthogonal (and hence LI) eigenvectors, and if these are normalized and used as the columns of $\mathbf{Q}$ then $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$, and

$$\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_n \end{bmatrix} \equiv \mathbf{D}, \tag{9}$$

where the λ_j's are the eigenvalues of $\mathbf{A}$, so (8) is the desired canonical form of f,

$$f = \lambda_1 \tilde{x}_1^2 + \cdots + \lambda_n \tilde{x}_n^2. \tag{10}$$

EXAMPLE 2. Reduce

$$f(x_1, x_2) = 3x_1^2 + 3x_2^2 + 2x_1x_2 \tag{11}$$

to canonical form. First, identify $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}. \tag{12}$$

*We do not claim that $\mathbf{Q}$ *must* be a modal matrix of $\mathbf{A}$ for $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$ to be diagonal.

Next, determine the eigenvalues and normalized eigenvectors of $\mathbf{A}$. These are found to be

$$\lambda_1 = 4,\ \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 2,\ \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{13}$$

(Whether we call $\lambda_1 = 4$ and $\lambda_2 = 2$, or vice versa, will not matter.) According to (10), the desired canonical form of f is then

$$f = \lambda_1 \tilde{x}_1^2 + \lambda_2 \tilde{x}_2^2 = 4\tilde{x}_1^2 + 2\tilde{x}_2^2. \tag{14}$$

It looks like we are done, so why do we need to know $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$? To know the connection between x_1, x_2 and $\tilde{x}_1, \tilde{x}_2$. Specifically,

$$\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2]\tilde{\mathbf{x}},$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}$$

so that

$$\begin{aligned} x_1 &= \frac{1}{\sqrt{2}}\tilde{x}_1 + \frac{1}{\sqrt{2}}\tilde{x}_2 \\ x_2 &= \frac{1}{\sqrt{2}}\tilde{x}_1 - \frac{1}{\sqrt{2}}\tilde{x}_2. \end{aligned} \tag{15}$$

Or, if we wish to have it the other way around, we can use $\tilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$. ▮

Let us summarize:

THEOREM 11.6.1 *Canonical Form*
A quadratic form $f(x_1, \ldots, x_n) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ can be reduced to the canonical form $\lambda_1\tilde{x}_1^2 + \cdots + \lambda_n\tilde{x}_n^2$ by introducing the change of variables $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where the λ_j's are the eigenvalues of $\mathbf{A}$ and the columns of $\mathbf{Q}$ are the corresponding normalized eigenvectors of $\mathbf{A}$. The reverse transformation is given by $\tilde{\mathbf{x}} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$.

A quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is classified as **positive definite** (i.e., "definitely positive") if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, and as **negative definite** if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$. Likewise, $\mathbf{A}$ is classified as positive (negative) definite if the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive (negative) definite.

THEOREM 11.6.2 *Definiteness*
Let $\mathbf{A}$ be symmetric. Then $\mathbf{A}$ and its quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ are positive (negative) definite if every eigenvalue of $\mathbf{A}$ is positive (negative).

Proof: It is simplest to work with the canonical form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \lambda_1\widetilde{x}_1^2 + \cdots + \lambda_n\widetilde{x}_n^2$. Since the latter is a sum of squares, it is evident that if $\lambda_1, \ldots, \lambda_n$ are all positive (negative), then $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive (negative) for all $\mathbf{x} \neq \mathbf{0}$. It remains to show that $\widetilde{\mathbf{x}} \neq \mathbf{0}$ if and only if $\mathbf{x} \neq \mathbf{0}$. Since $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$ (i.e., $\mathbf{Q}$ is nonsingular), $\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{Q}\mathbf{0} = \mathbf{0}$, and $\mathbf{x} = \mathbf{0}$ implies that $\widetilde{\mathbf{x}} = \mathbf{Q}^{\mathrm{T}}\mathbf{0} = \mathbf{0}$. ■

For instance, f and $\mathbf{A}$ in Example 2 are positive definite because $\lambda_1 = 4 > 0$ and $\lambda_2 = 2 > 0$.

EXAMPLE 3. If

$$f(x_1, x_2, x_3) = 2x_1^2 + 4x_2^2 + 4x_3^2 + 2x_1x_2 - 2x_1x_3 + 6x_2x_3, \tag{16}$$

then

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix} \quad \text{and} \quad \lambda_1 = 7,\ \lambda_2 = 3,\ \lambda_3 = 0. \tag{17}$$

It might appear that f is positive definite but it is not, because λ_3 is not positive. Being zero is not good enough, for remember that being positive definite means that $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ or, equivalently, $\widetilde{\mathbf{x}}^{\mathrm{T}}(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})\widetilde{\mathbf{x}} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. Yet,

$$f = 7\widetilde{x}_1^2 + 3\widetilde{x}_2^2 + 0\widetilde{x}_3^2 \tag{18}$$

is zero if

$$\widetilde{\mathbf{x}} = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \\ \widetilde{x}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 8 \end{bmatrix}, \tag{19}$$

for instance. ■

Let us conclude this section with a physical application.

EXAMPLE 4. *Buckling Load.* Three rigid rods, each of length L, are pinned at their ends, with the end A constrained to move in a frictionless vertical slot, as shown in Fig. 1. The three lateral springs, each of stiffness k, are unstretched and uncompressed when the system is undeflected (i.e., when $x = y = 0$), and the middle spring is attached at the middle of the middle rod. We find that as the load P is slowly increased the deflection remains zero ($x = y = z = 0$) so that $x = y = z = 0$ constitutes an equilibrium state of the system. Eventually, however, when P reaches a critical value, say P_{cr}, the system "buckles," that is, collapses. The problem posed here is to determine P_{cr} in terms of the given quantities. It will be convenient to work with the potential energy V of the system, and to use the physical principle that the system will arrange itself so as to minimize its potential V.

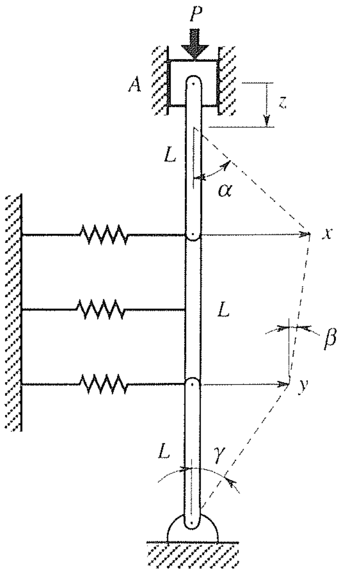


Figure 1. Buckling.

Recall, from elementary physics, that the potential energy stored in a spring of stiffness k, deflected by an amount x, is $\frac{1}{2}kx^2$. Further, note that we may associate a potential with the load P by regarding it as being due to a weight P sitting on the slider. Then the gravitational potential of the weight is $-Pz$, if we define the potential to be zero when $z = 0$. Thus,

$$V = \frac{1}{2}kx^2 + \frac{1}{2}ky^2 + \frac{1}{2}k\left(\frac{x+y}{2}\right)^2 - Pz. \tag{20}$$

But x, y, z are not independent variables. For instance, we can express z in terms of x and y through the geometry. Specifically, if we introduce the angles α, β, γ, shown in Fig. 1, then

$$z + L\cos\alpha + L\cos\beta + L\cos\gamma = 3L. \tag{21}$$

For the onset of the buckling x, y, z are all infinitesimal (arbitrarily small), so

$$\begin{aligned}\cos\alpha &= \sqrt{1-\sin^2\alpha} = \sqrt{1-\left(\frac{x}{L}\right)^2} \\ &= 1 - \frac{1}{2}\left(\frac{x}{L}\right)^2 - \frac{1}{8}\left(\frac{x}{L}\right)^4 - \cdots \quad \text{(Taylor series)} \\ &\sim 1 - \frac{1}{2}\left(\frac{x}{L}\right)^2 \quad \text{as } x \to 0.\end{aligned} \tag{22}$$

Similarly for $\cos\beta$ and $\cos\gamma$ so (21) becomes

$$z + L\left[1 - \frac{1}{2}\left(\frac{x}{L}\right)^2\right] + L\left[1 - \frac{1}{2}\left(\frac{x-y}{L}\right)^2\right] + L\left[1 - \frac{1}{2}\left(\frac{y}{L}\right)^2\right] \sim 3L \tag{23}$$

or, upon simplification,

$$z \sim \frac{x^2 - xy + y^2}{L} \tag{24}$$

so that

$$V(x,y) \sim \left(\frac{5k}{8} - \frac{P}{L}\right)x^2 + \left(\frac{5k}{8} - \frac{P}{L}\right)y^2 + \left(\frac{k}{4} + \frac{P}{L}\right)xy. \tag{25}$$

The right-hand side of (25) is a quadratic form

$$f(x,y) = \left(\frac{5k}{8} - \frac{P}{L}\right)x^2 + \left(\frac{5k}{8} - \frac{P}{L}\right)y^2 + \left(\frac{k}{4} + \frac{P}{L}\right)xy, \tag{26}$$

with the associated matrix

$$\mathbf{A} = \begin{bmatrix} \frac{5k}{8} - \frac{P}{L} & \frac{k}{8} + \frac{P}{2L} \\ \frac{k}{8} + \frac{P}{2L} & \frac{5k}{8} - \frac{P}{L} \end{bmatrix}. \tag{27}$$

The crucial point is whether or not f is positive definite for if $f(x,y)$ is positive definite, then $V(x,y)$ has a minimum at $x = y = 0$ and the undeflected equilibrium configuration $x = y = 0$ is stable. Otherwise, the equilibrium configuration will be unstable and the system will buckle.

To assess the positive definiteness of f, we need merely evaluate the eigenvalues of $\mathbf{A}$. These are found to be

$$\lambda_1 = \frac{3k}{2}\left(\frac{1}{3} - \frac{P}{kL}\right), \qquad \lambda_2 = \frac{k}{2}\left(\frac{3}{2} - \frac{P}{kL}\right) \tag{28}$$

from which we see that f is positive definite (i.e., both λ_j's are positive) if and only if $P/(kL) < \frac{1}{3}$. Thus, the critical load, the "buckling load," is $P_{cr} = kL/3$.

COMMENT 1. If $P/(kL) < \frac{1}{3}$, then both λ_j's are positive in the canonical version $\lambda_1 \widetilde{x}^2 + \lambda_2 \widetilde{y}^2$ of V so the graph of V has a "valley" at the origin ($\widetilde{x} = \widetilde{y} = 0$ or, equivalently $x = y = 0$). If $P/(kL) > \frac{3}{2}$, then both λ_j's are negative so the graph of V has a "hill" at the origin. And if $\frac{1}{3} < P/(kL) < \frac{3}{2}$, then $\lambda_1 < 0$ and $\lambda_2 > 0$ so the graph of V has a "saddle" at the origin (i.e., like an equestrian saddle it has a valley in one direction and a hill in the other). In the case of the valley V has a *minimum* at the origin so the equilibrium solution $x = y = 0$ is *stable* and buckling does not occur. In the other two cases (hill and saddle) V does not have a minimum at the origin, so the equilibrium solution $x = y = 0$ is unstable and buckling occurs. The borderline between these two cases gives the critical buckling load.

COMMENT 2. Physically, the idea is that as the system begins to deflect, the springs gain potential energy while the weight P loses potential energy. For instability (buckling) we need the loss to exceed the gain. Since the loss is proportional to P, one anticipates the existence of a critical value P_{cr} such that if $P = P_{cr}$, the loss just balances the gain, and if $P > P_{cr}$, then the loss exceeds the gain. ▮

Closure. The chief result of this section is given in Theorem 11.6.1, that a quadratic form can be reduced to canonical form by a normalized modal matrix. In addition, Theorem 11.6.2 tells us that a quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive definite if every eigenvalue of $\mathbf{A}$ is positive, and negative definite if every eigenvalue is negative. Positive and negative definiteness is central when we discuss the maxima and minima of functions of several variables in Chapter 13.

EXERCISES 11.6

1. For each of the following quadratic forms in n variables determine the associated symmetric $\mathbf{A}$ matrix.

(a) $2x_1^2 + 4x_2^2 + x_1x_2 \quad (n = 2)$
(b) $x_3^2 + 2x_1x_3 + 2x_2x_3 \quad (n = 3)$
(c) $x_1^2 + x_2^2 + x_1x_2 \quad (n = 2)$
(d) $x_1^2 - 4x_3^2 + 3x_1x_2 \quad (n = 3)$
(e) $x_1^2 + x_4^2 + 2x_1x_4 \quad (n = 4)$
(f) $4x_1x_4 + 4x_2x_3 \quad (n = 4)$
(g) $3x_1x_2 \quad (n = 2)$
(h) $4x_1x_2 - 2x_3^2 \quad (n = 3)$
(i) $-2x_1x_3 \quad (n = 3)$
(j) $x_3^2 + x_1x_2 + x_1x_3 + x_2x_3 \quad (n = 3)$
(k) $6x_2x_3 \quad (n = 3)$
(l) $2x_1x_2 + 2x_1x_3 + 2x_2x_3 - 4x_4^2 \quad (n = 4)$

2. (a)–(l) Reduce each quadratic form in Exercise 1 to the canonical form (10) by means of a normalized modal matrix transformation $\mathbf{Q}$. Further, classify the quadratic form and the associated symmetric $\mathbf{A}$ matrix as positive definite or negative definite, where applicable

3. We state, without proof, that a *necessary and sufficient con-*

dition for a quadratic form $f = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, *and its associated symmetric matrix* $\mathbf{A}$, *to be positive definite is that*

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0,$$

$$\ldots, \quad \det\mathbf{A} > 0.$$

Apply this condition to each of the following $\mathbf{A}$ matrices, and compare your conclusion with that drawn from direct examination of the eigenvalues.

(a) $\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} -2 & 1 & 2 \\ 1 & -2 & 2 \\ 2 & 2 & 1 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 9 \end{bmatrix}$

4. Reduce the following quadratic equations to the canonical form $a\tilde{x}^2 + b\tilde{y}^2 = c$, and hence decide whether they correspond to *ellipses* or *hyperbolas*. Sketch their graphs, showing both the x, y and $\tilde{x}, \tilde{y}$ axes, as well as the intercepts on the $\tilde{x}, \tilde{y}$ axes. HINT: Recall that $(\tilde{x}/A)^2 + (\tilde{y}/B)^2 = 1$ is the equation of an ellipse with intercepts at $\pm A$ on the $\tilde{x}$ axis and $\pm B$ on the $\tilde{y}$ axis (or a circle of radius A if $A = B$). Further, $(\tilde{x}/A)^2 - (\tilde{y}/B)^2 = 1$ is the equation of a hyperbola with intercepts at $\pm A$ on the $\tilde{x}$ axis, and $(\tilde{y}/B)^2 - (\tilde{x}/A)^2 = 1$ is the equation of a hyperbola with intercepts at $\pm B$ on the $\tilde{y}$ axis.

(a) $3x^2 + 2y^2 - 2xy = 6$
(b) $xy = 1$
(c) $4y^2 + 3xy = 1$
(d) $x^2 + y^2 - 10xy = 4$
(e) $x^2 + y^2 + 2xy = 4$ (This case will be found to correspond to a limiting case. Explain what we mean by that.)

5. (*Completing squares*) The successful $\mathbf{Q}$ matrix, in (7), is not necessarily a modal matrix of $\mathbf{A}$. Given $f = x_1^2 + x_2^2 + x_1x_2$, for example, suppose that we proceed, instead, by "completing squares:"

$$\begin{aligned} f &= x_1^2 + x_2^2 + x_1x_2 \\ &= \left(x_1^2 + x_1x_2 + \tfrac{1}{4}x_2^2\right) + x_2^2 - \tfrac{1}{4}x_2^2 \\ &= \left(x_1 + \tfrac{1}{2}x_2\right)^2 + \tfrac{3}{4}x_2^2, \end{aligned} \tag{5.1}$$

so the transformation

$$\tilde{x}_1 = x_1 + \tfrac{1}{2}x_2, \quad \tilde{x}_2 = x_2 \tag{5.2}$$

reduces f to the canonical form $f = \tilde{x}_1^2 + \frac{3}{4}\tilde{x}_2^2$.

(a) Show that the matrix $\mathbf{Q}$ corresponding to (5.2) is *not* a modal matrix.
(b) Reduce $f = x_1^2 + x_2^2 + 4x_3^2 + 2x_1x_2 - x_1x_3$ to canonical form by completing squares.
(c) Repeat part (b), for $f = x_1^2 + 4x_1x_2 + x_2x_3$.
(d) Repeat part (b), for $f = 4x_1^2 + 2x_1x_3 + x_2x_3$.
(e) Repeat part (b), for $f = x_3^2 + 2x_1x_3$.
(f) Does the method of completing squares work for $f = 2x_1x_2$? Explain.

6. Rework Example 4 for the case where the middle spring is removed. Show that the buckling load is then $P_{cr} = \frac{1}{3}kL$. Note that this result is the same as in Example 4, where the middle spring is included. Explain, in physical terms, why this is so.

Chapter 11 Review

The matrix eigenvalue problem is the search for nontrivial solutions of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ or, equivalently,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}; \tag{1}$$

that is, solutions other than the trivial solution $\mathbf{x} = \mathbf{0}$. The eigenvalues are found by setting

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0, \tag{2}$$

which condition guarantees the existence of nontrivial solutions of (1). Known as the characteristic equation of $\mathbf{A}$, (2) is an nth-degree polynomial equation, which always has at least one and at most n distinct roots. For each eigenvalue λ_j thus found, the solution of $(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{x} = \mathbf{0}$ by Gauss elimination gives the corresponding eigenvectors $\mathbf{e}_j$.

Rather than being rare, the case of symmetric matrices is common in applications (as noted, for instance, in Example 2 of Section 11.3, Example 1 in Section 11.4, and in all of Section 11.6). If an $n \times n$ matrix $\mathbf{A}$ is symmetric, then all of its eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and its eigenvectors provide an orthogonal basis for n-space.

A pattern emerges insofar as choice of basis. Namely, when a basis is needed, to expand vectors, the most convenient basis to use is probably the basis provided by the eigenvectors of the $\mathbf{A}$ matrix to be found within the given problem, For instance, we do that to study the stability of the equilibrium solution of the Markov process (Example 4, Section 11.2), to solve the nonhomogeneous equation $\mathbf{Ax} = \lambda\mathbf{x} + \mathbf{c}$ (Section 11.3.2), and to prove the convergence of the power method (Exercise 12, Section 11.3). The reason that eigenvector bases are convenient is that if we multiply a vector equation by $\mathbf{A}$, then we need to evaluate the vectors $\mathbf{Ae}_j$. If the $\mathbf{e}_j$'s are eigenvectors of $\mathbf{A}$, then $\mathbf{Ae}_j$ is simply the single term $\lambda_j \mathbf{e}_j$; if not, it is a linear combination of the n base vectors, $\mathbf{e}_1, \dots, \mathbf{e}_n$.

In Section 11.4 we study the diagonalization of an $n \times n$ matrix $\mathbf{A}$. There, Theorem 11.4.1 is the most significant result because it gives a necessary and sufficient condition for $\mathbf{A}$ to be diagonalizable (namely, that it have n LI eigenvectors), and it tells us how to choose $\mathbf{Q}$ so that

$$\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{D} \tag{3}$$

is diagonal. Namely, if we use the (n LI) eigenvectors of $\mathbf{A}$ as the columns of $\mathbf{Q}$, then $\mathbf{D} = \{d_{ij}\}$ is diagonal, with $d_{jj} = \lambda_j$. Further, Theorem 11.4.3 gives a sufficient condition for diagonalizability, that $\mathbf{A}$ have n distinct eigenvalues. Since the generic case is for the characteristic equation to have distinct roots, the generic case is for a given $n \times n$ matrix to be diagonalizable. If $\mathbf{A}$ is symmetric, then it is diagonalizable whether or *not* it has n distinct eigenvalues because every $n \times n$ symmetric matrix has n orthogonal (and hence LI) eigenvectors. For symmetric matrices we urge you to use the *normalized* eigenvectors of $\mathbf{A}$ to form its modal matrix $\mathbf{Q}$ because then $\mathbf{Q}$ admits the useful property that

$$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}, \tag{4}$$

that is, the inverse of $\mathbf{Q}$ is simply its transpose.

Even if $\mathbf{A}$ is not diagonalizable, it can be reduced to Jordan normal form. That is, a generalized modal matrix $\mathbf{P}$ can be found so that

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{J} \tag{5}$$

is triangular; $\mathbf{J}$ is upper triangular with zeros above its main diagonal except for 1's immediately above one or more diagonal elements. That case is left for the exercises.

Whereas in Section 11.4 we concentrate on the concept of diagonalization, and show how to use diagonalization to uncouple systems of differential equations, in Section 11.5 we shift our focus to differential equation theory itself, using matrix theory and diagonalization as tools, and solve the general system of n first-order nonhomogeneous linear differential equations with constant coefficients.

In the final section, 11.6, we show that a quadratic form can be expressed in matrix terminology as $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is symmetric, and can be reduced to canonical form $\lambda_1\widetilde{x}_1^2 + \cdots + \lambda_n\widetilde{x}_n^2$ by the change of variables $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$, where $\mathbf{Q}$ is a normalized modal matrix of $\mathbf{A}$. That is, $\mathbf{Q} = [\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_n]$, where $\hat{\mathbf{e}}_j$ is a normalized eigenvector corresponding to the eigenvalue λ_j.

Applications of the eigenvalue problem are extensive. Examples studied in this chapter are drawn from the areas of oscillation theory, systems of coupled differential equations, and population dynamics, with other applications covered in the exercises. In Chapter 13 we use the theory of quadratic forms to help us classify the extrema of functions of several variables.

Chapter 12

Extension to Complex Case (Optional)

PREREQUISITE: Familiarity with the algebra of complex numbers, covered in Section 21.2.

12.1 Introduction

In Chapters 8–11 all scalars are understood to be real: the coefficients in systems of linear equations, the components of vectors, the elements of matrices, and so on. In some applications, however, perhaps more so in physics and chemistry than in engineering, complex numbers enter. For example, you may have met the *Euler angles* θ, ϕ, ψ used to specify the orientation of a rigid body such as a gyroscope. Alternatively, it is sometimes advantageous to employ so-called *Cayley–Klein parameters*, and in doing so one meets the *Pauli spin matrices*

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{1}$$

of which σ_y is seen to have complex elements.

The purpose of this chapter is to indicate the changes that need to be made in extending our vector and matrix systems so as to include complex numbers. Although this could have been done from the start, it was felt that the gain in simplicity, in the preceding chapters, offsets the need for this special chapter, which we make quite brief.

12.2 Complex n-Space

All scalars in the vector space $\mathbb{R}^n$ defined in Section 9.4 (namely, the scalars that multiplied vectors and the scalar components of the vectors themselves) are real.

If we allow these scalars to be complex, then in place of $\mathbb{R}^n$ we have the **complex n-space** denoted here as $\mathbb{C}^n$:

$$\mathbb{C}^n = \{(a_1, \dots, a_n) \mid a_1, \dots, a_n \text{ complex numbers } \}. \tag{1}$$

The definitions $\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, \dots, u_n + v_n)$, $\alpha\mathbf{u} \equiv (\alpha u_1, \dots, \alpha u_n)$, $\mathbf{0} \equiv (0, \dots, 0)$, and $-\mathbf{u} \equiv (-u_1, \dots, -u_n)$ are the same as for $\mathbb{R}^n$, except that now the scalars are complex numbers. From these definitions the same properties follow as for $\mathbb{R}^n$ ($\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$, etc.) as in (10) in Section 9.4.

However, the Euclidean dot product $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$, which we adopted for real vector space, is unacceptable because the resulting norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{u_1^2 + \cdots + u_n^2}$ fails to display the key properties expected of a norm, in particular, the nonnegativeness condition

$$\begin{aligned} \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}. \end{aligned} \tag{2}$$

For example, if $\mathbf{u} = [2, 2i, 0, 0]$, then $\|\mathbf{u}\| = \sqrt{(2)^2 + (2i)^2 + (0)^2 + (0)^2} = 0$ even though $\mathbf{u} \neq \mathbf{0}$; and if $\mathbf{u} = [0, 2i, 0, 0]$, then $\|\mathbf{u}\| = \sqrt{0 - 4 + 0 + 0} = 2i$ is not even real so it cannot satisfy the condition $\|\mathbf{u}\| > 0$.*

To avoid this problem with the norm, we adopt the modified dot product

$$\boxed{\mathbf{u} \cdot \mathbf{v} \equiv u_1 \overline{v}_1 + u_2 \overline{v}_2 + \cdots + u_n \overline{v}_n = \sum_{j=1}^{n} u_j \overline{v}_j} \tag{3}$$

where the overhead bar denotes complex conjugate for then

$$\boxed{\|\mathbf{u}\| \equiv \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j \overline{u}_j} = \sqrt{\sum_{j=1}^{n} |u_j|^2}} \tag{4}$$

does satisfy the nonnegativeness condition (2). [Recall that if $z = a + ib$, then $z\overline{z} = (a + ib)(a - ib) = a^2 + b^2 = |z|^2$.]

EXAMPLE 1. If $\mathbf{u} = [2, 3 - 5i, 0, 4i]$, then

$$\begin{aligned} \|\mathbf{u}\| &= \sqrt{(2)(2) + (3 - 5i)(3 + 5i) + (0)(0) + (4i)(-4i)} \\ &= \sqrt{4 + 34 + 16} = \sqrt{54}, \end{aligned}$$

*Recall from Section 22.2 (which is the prerequisite for this chapter) that inequalities such as $z > 0$ and $z < 0$ are not meaningful if z is complex; see the paragraph below equation (12) in that section.

which *is* real and positive. ▮

Properties of the dot product. Observe that

$$\begin{aligned}
\overline{\mathbf{v}\cdot\mathbf{u}} &= \overline{v_1\overline{u}_1 + \cdots + v_n\overline{u}_n} \\
&= \overline{v_1\overline{u}_1} + \cdots + \overline{v_n\overline{u}_n} \\
&= \overline{v}_1\overline{\overline{u}}_1 + \cdots + \overline{v}_n\overline{\overline{u}}_n \\
&= \overline{v}_1 u_1 + \cdots + \overline{v}_n u_n \\
&= u_1\overline{v}_1 + \cdots + u_n\overline{v}_n \\
&= \mathbf{u}\cdot\mathbf{v}.
\end{aligned}$$

Thus, in place of the commutativity condition $\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$, satisfied in the real case, we have the so-called conjugate commutativity $\mathbf{u}\cdot\mathbf{v} = \overline{\mathbf{v}\cdot\mathbf{u}}$ in the complex case. In fact, the properties

$$\begin{aligned}
&\textit{Conjugate Commutative:} & \mathbf{u}\cdot\mathbf{v} &= \overline{\mathbf{v}\cdot\mathbf{u}}, && \text{(5a)} \\
&\textit{Nonnegative:} & \mathbf{u}\cdot\mathbf{u} &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, && \\
& & &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, && \text{(5b)} \\
&\textit{Linear:} & (\alpha\mathbf{u} + \beta\mathbf{v})\cdot\mathbf{w} &= \alpha(\mathbf{u}\cdot\mathbf{w}) + \beta(\mathbf{v}\cdot\mathbf{w}), && \text{(5c)}
\end{aligned}$$

of the complex dot product are the same as in the real case [(12) in Section 9.5.2] except for the complex conjugate bar in (5a).

EXAMPLE 2. If $\mathbf{u} = [1 + 2i, -4]$ and $\mathbf{v} = [i, 3 - i]$, then $\mathbf{u}\cdot\mathbf{v} = (1 + 2i)(-i) + (-4)(3 + i) = -10 - 5i$, and $\mathbf{v}\cdot\mathbf{u} = (i)(1 - 2i) + (3 - i)(-4) = -10 + 5i$, which does equal $\overline{\mathbf{u}\cdot\mathbf{v}}$, in accordance with (5a). ▮

The **Schwarz inequality,**

$$|\mathbf{u}\cdot\mathbf{v}| \leq \|\mathbf{u}\|\,\|\mathbf{v}\|, \tag{6}$$

is found to hold (Exercise 7), as in the real case, except that here $|\mathbf{u}\cdot\mathbf{v}|$ is the modulus of a complex number rather than the absolute value of a real number.

Properties of the norm. The norm (4) admits the properties

$$\begin{aligned}
&\textit{Scaling:} & \|\alpha\mathbf{u}\| &= |\alpha|\,\|\mathbf{u}\|, && \text{(7a)} \\
&\textit{Nonnegative:} & \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, && \\
& & &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}, && \text{(7b)} \\
&\textit{Triangle Inequality:} & \|\mathbf{u} + \mathbf{v}\| &\leq \|\mathbf{u}\| + \|\mathbf{v}\|. && \text{(7c)}
\end{aligned}$$

These properties are identical to those for the real case [(17) in Section 9.5.3] but here $|\alpha|$ is the modulus of a complex number rather than the absolute value of a real number. Also, the proofs are slightly different due to the presence of complex numbers. To illustrate, consider (7c):

$$\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) \\
&= \mathbf{u}\cdot\mathbf{u}+\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{v} \\
&= \|\mathbf{u}\|^2+\mathbf{u}\cdot\mathbf{v}+\overline{\mathbf{u}\cdot\mathbf{v}}+\|\mathbf{v}\|^2 \\
&= \|\mathbf{u}\|^2+2\,\mathrm{Re}(\mathbf{u}\cdot\mathbf{v})+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,|\mathrm{Re}(\mathbf{u}\cdot\mathbf{v})|+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,|\mathbf{u}\cdot\mathbf{v}|+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,\|\mathbf{u}\|\,\|\mathbf{v}\|+\|\mathbf{v}\|^2 \\
&= (\|\mathbf{u}\|+\|\mathbf{v}\|)^2
\end{aligned}$$

so that

$$\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\|+\|\mathbf{v}\|.$$

Proofs of (7a) and (7b) are left for the exercises.

Thus far, then, the only substantial change (beyond the fact that the scalars are now complex) is the change in the dot product from $\mathbf{u}\cdot\mathbf{v} \equiv u_1v_1+\cdots+u_nv_n$ to $\mathbf{u}\cdot\mathbf{v} \equiv u_1\overline{v}_1+\cdots+u_n\overline{v}_n$. Furthermore, whereas we defined the angle θ between $\mathbf{u}$ and $\mathbf{v}$ as $\theta = \cos^{-1}(\mathbf{u}\cdot\mathbf{v}/\|\mathbf{u}\|\,\|\mathbf{v}\|)$ in the real case, this definition is awkward in the complex case since $\mathbf{u}\cdot\mathbf{v}$ (and hence the argument of the arccosine) is complex. Instead of trying to patch things up, we simply choose not to define θ for the complex case, although we do retain a notion of perpendicularity; that is, we still say that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\mathbf{u}\cdot\mathbf{v}=0$.

Finally,

THEOREM 12.2.1 *Dimension of* $\mathbb{C}^n$
The dimension of $\mathbb{C}^n$ is n.

You may have expected the dimension to be $2n$, on the grounds that each of the n vector components has both real and imaginary parts. However, observe that the ON set

$$\begin{aligned}
\hat{\mathbf{e}}_1 &= [1,0,\dots,0], \\
\hat{\mathbf{e}}_2 &= [0,1,0,\dots,0], \\
&\vdots \\
\hat{\mathbf{e}}_n &= [0,\dots,0,1]
\end{aligned} \tag{8}$$

is a basis for $\mathbb{C}^n$, just as it is for $\mathbb{R}^n$ since the set is LI and spans the space. That it spans $\mathbb{C}^n$ follows from the fact that every vector $\mathbf{u}=(u_1,\dots,u_n)$ in $\mathbb{C}^n$ can be expanded as $\mathbf{u} = u_1\hat{\mathbf{e}}_1 + \cdots + u_n\hat{\mathbf{e}}_n$. And since the basis (8) contains n vectors, it follows from the definition of dimension in Section 9.9.2 that $\mathbb{C}^n$ is n-dimensional.

EXERCISES 12.2

1. Normalize each of the following vectors

(a) $[1, i]$ (b) $[1+i, 1-i]$
(c) $[1, 3, -2, 0]$ (d) $[2, 1-3i, 0, 5]$
(e) $[2+3i, 1-i, 4i]$ (f) $[i, 0, 0, -i]$
(g) $[x, y, z, ict]$ This vector (in which x, y, z are Cartesian coordinates, c is the speed of light, and t is the time) arises in the *special theory of relativity*. In that application, the space $\mathbb{C}^4$ is known as *world space*, or *Minkowski space*.

2. Show whether the following vector sets are bases for $\mathbb{C}^3$.

(a) $[i, 2, 1+i]$, $[0, 1, 2+i]$, $[4, 1, -i]$
(b) $[1, 0, 2]$, $[3, 2, -2]$, $[0, 0, 4]$
(c) $[4, 1-2i, 0]$, $[3+i, i, -2i]$, $[-2-2i, 1-4i, 4i]$
(d) $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, i]$
(e) $[i, 1, 2]$, $[3, 1-i, -i]$, $[3+2i, 3-i, 4-i]$, $[3-i, -i, -2-i]$

3. Show that the set $\mathbf{e}_1 = [i, 1, 0]$, $\mathbf{e}_2 = [2, 2i, 1]$, $\mathbf{e}_3 = [1, i, -4]$ is an orthogonal basis for $\mathbb{C}^3$.

4. Show that (23) in Section 9.9 (with $k = n$) is valid for $\mathbb{C}^n$, and use it to expand each vector in terms of the bases given in Exercise 3.

(a) $[2+4i, -3i, 2]$ (b) $[0, 0, 1]$
(c) $[1, 1, 1]$ (d) $[i, 2i, 3i]$
(e) $[0, i, 0]$ (f) $[1-i, 0, 0]$

5. Show that

$$(\alpha\mathbf{x}) \cdot \mathbf{y} = \alpha(\mathbf{x} \cdot \mathbf{y}), \tag{5.1}$$

whereas

$$\mathbf{x} \cdot (\alpha\mathbf{y}) = \overline{\alpha}(\mathbf{x} \cdot \mathbf{y}). \tag{5.2}$$

6. The property (5a) was proved in the text. Prove (5b) and (5c).

7. Prove the Schwarz inequality (6).

8. Prove the properties (7a) and (7b).

9. Vectors in $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$ can be displayed, graphically, as arrow vectors. Is the same true for $\mathbb{C}^1$, $\mathbb{C}^2$, $\mathbb{C}^3$? Explain.

12.3 Complex Matrices

All of Chapter 10 (on matrices, determinants, and linear equations) holds even if the scalars are allowed to be complex. To illustrate, consider a representative example.

EXAMPLE 1. Given

$$\mathbf{A} = \begin{bmatrix} 2 & 1+i \\ 0 & i \end{bmatrix}, \tag{1}$$

compute $\mathbf{A}^{-1}$. We proceed as usual, although the numbers are now complex:

$$\det\mathbf{A} = (2)(i) - (0)(1+i) = 2i \quad (\neq 0, \text{ so } \mathbf{A}^{-1} \text{ exists})$$

$$\begin{array}{llll} \text{Minors:} & M_{11} = i & \text{Cofactors:} & A_{11} = i \\ & M_{12} = 0 & & A_{12} = 0 \\ & M_{21} = 1+i & & A_{21} = -1-i \\ & M_{22} = 2 & & A_{22} = 2 \end{array}$$

so

$$\mathbf{A}^{-1} = \frac{1}{2i}\begin{bmatrix} i & -1-i \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{-1+i}{2} \\ 0 & -i \end{bmatrix}, \tag{2}$$

as is verified by showing that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ▮

However, in Chapter 11, on the eigenvalue problem, dot products begin to appear so the impact of the change in the dot product (from $\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n$ to $\mathbf{x} \cdot \mathbf{y} = x_1 \overline{y}_1 + \cdots + x_n \overline{y}_n$) begins to be felt. Expressed in matrix form, the dot product is now

$$\boxed{\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{y}}} \quad \text{or} \quad \overline{\mathbf{y}}^{\mathrm{T}}\mathbf{x}. \tag{3}$$

We begin with two definitions. Given an $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, we define the **complex conjugate** of $\mathbf{A}$ as

$$\overline{\mathbf{A}} \equiv \{\overline{a}_{ij}\} \tag{4}$$

and the **Hermitian conjugate** of $\mathbf{A}$ as

$$\mathbf{A}^* \equiv \{\overline{a}_{ji}\}, \quad \text{i.e.,} \quad \boxed{\mathbf{A}^* \equiv \overline{\mathbf{A}}^{\mathrm{T}}.} \tag{5}$$

If $\overline{\mathbf{A}} = \mathbf{A}$ then $\mathbf{A}$ is **real**, and if $\mathbf{A}^* = \mathbf{A}$ then $\mathbf{A}$ is **Hermitian.**[†] If $\mathbf{A}$ is not square then it cannot be Hermitian.

EXAMPLE 2. If

$$\mathbf{A} = \begin{bmatrix} 2+i & 0 & 3-5i \\ 7 & 1 & 4i \\ 2 & i & 3 \end{bmatrix},$$

then

$$\overline{\mathbf{A}} = \begin{bmatrix} 2-i & 0 & 3+5i \\ 7 & 1 & -4i \\ 2 & -i & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^* = \begin{bmatrix} 2-i & 7 & 2 \\ 0 & 1 & -i \\ 3+5i & -4i & 3 \end{bmatrix}.$$

Since $\mathbf{A}^* \neq \mathbf{A}$, $\mathbf{A}$ is not Hermitian. ▮

EXAMPLE 3. If

$$\mathbf{A} = \begin{bmatrix} 3 & 1+4i \\ 1-4i & 0 \end{bmatrix},$$

[†] *Charles Hermite* (1822–1901), a professor at the Sorbonne and at the Ecole Polytechnique, contributed to the theory of elliptic functions and is also well known for his introduction of the Hermite polynomials.

then

$$\overline{\mathbf{A}} = \begin{bmatrix} 3 & 1-4i \\ 1+4i & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^* = \begin{bmatrix} 3 & 1+4i \\ 1-4i & 0 \end{bmatrix}.$$

Since $\mathbf{A}^* = \mathbf{A}$, $\mathbf{A}$ is Hermitian. ▮

Some properties of the complex conjugate and Hermitian conjugate matrices are as follows:

$$\overline{\overline{\mathbf{A}}} = \mathbf{A}, \tag{6a}$$

$$\overline{\mathbf{A}+\mathbf{B}} = \overline{\mathbf{A}} + \overline{\mathbf{B}}, \tag{6b}$$

$$\overline{\mathbf{A}\mathbf{B}} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}, \tag{6c}$$

$$(\mathbf{A}^*)^* = \mathbf{A}, \tag{6d}$$

$$(\mathbf{A}+\mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*, \tag{6e}$$

$$(\mathbf{A}\mathbf{B})^* = \mathbf{B}^*\mathbf{A}^*. \tag{6f}$$

These properties are readily verified. In addition, the key property of the Hermitian conjugate matrix $\mathbf{A}^*$ is that

$$\boxed{(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y})} \tag{7}$$

holds for all vectors $\mathbf{x}$ and $\mathbf{y}$; specifically, if $\mathbf{A}$ is $m \times n$, $\mathbf{x}$ is any $n \times 1$ vector and $\mathbf{y}$ is any $m \times 1$ vector. To prove (7), observe that

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = (\mathbf{A}\mathbf{x})^{\mathrm{T}}\overline{\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\overline{\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\overline{\overline{\mathbf{A}}^{\mathrm{T}}\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{A}^*\mathbf{y}} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y}). \tag{8}$$

Hermitian matrices. Recall from Chapters 10 and 11 that matrices arising in applications are often *symmetric* ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$) and that such matrices exhibit several useful properties concerning the eigenvalue problem (Theorems 11.3.1–11.3.4). Likewise, when complex matrices arise in applications they are often *Hermitian* ($\mathbf{A}^* = \overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$), and such matrices exhibit analogous useful properties, given by Theorems 12.3.1–12.3.4 below.

EXAMPLE 4. *Lorentz Transformation.* In the special theory of relativity one considers the vector $[x, y, z, ict]$, where x, y, z are Cartesian coordinates, c is the speed of light, and t is the time. If the corresponding vector, referred to an x', y', z' system which is translating in the z direction with constant speed v, is denoted as $[x', y', z', ict']$, it turns out that these vectors are related according to

$$\begin{bmatrix} x' \\ y' \\ z' \\ ict' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \dfrac{1}{\sqrt{1-\beta^2}} & \dfrac{i\beta}{\sqrt{1-\beta^2}} \\ 0 & 0 & \dfrac{-i\beta}{\sqrt{1-\beta^2}} & \dfrac{1}{\sqrt{1-\beta^2}} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ ict \end{bmatrix}, \tag{9}$$

where $\beta = v/c\,(< 1)$. This is the important *Lorentz transformation*, and it is seen that the transformation matrix is Hermitian. ▮

THEOREM 12.3.1 *Real Eigenvalues*
If $\mathbf{A}$ is Hermitian ($\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$), then all of its eigenvalues are real.

Proof: Let us both pre-dot and post-dot both sides of $\mathbf{Ax} = \lambda\mathbf{x}$ with $\mathbf{x}$:

$$\begin{aligned} \text{Pre-dot:} \quad \mathbf{x}\cdot(\mathbf{Ax}) &= \mathbf{x}\cdot(\lambda\mathbf{x}) \\ &= \overline{\lambda}(\mathbf{x}\cdot\mathbf{x}) \quad \text{[according to (3)]} \qquad (10a) \\ \text{Post-dot:} \quad (\mathbf{Ax})\cdot\mathbf{x} &= (\lambda\mathbf{x})\cdot\mathbf{x} \\ &= \lambda(\mathbf{x}\cdot\mathbf{x}) \quad \text{[according to (3)].} \qquad (10b) \end{aligned}$$

But the left-hand sides are equal, by virtue of property (7) together with the assumption that $\mathbf{A}$ is Hermitian. Thus, subtracting (10b) from (10a) gives

$$(\overline{\lambda} - \lambda)(\mathbf{x}\cdot\mathbf{x}) = 0. \qquad (11)$$

Now, $\mathbf{x}\cdot\mathbf{x} = \|\mathbf{x}\|^2 \neq 0$ since $\mathbf{x}$ is an eigenvector, so it follows from (11) that $\overline{\lambda} - \lambda = 0$, or $\overline{\lambda} = \lambda$. Thus, λ is real. ▪

THEOREM 12.3.2 *Dimension of Eigenspace*
If an eigenvalue λ of an Hermitian matrix $\mathbf{A}$ is of multiplicity k, then the eigenspace corresponding to λ is of dimension k.

THEOREM 12.3.3 *Orthogonality of Eigenvectors*
If $\mathbf{A}$ is Hermitian, then eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof: Let $\mathbf{e}_j$ and $\mathbf{e}_k$ be eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively. That is,

$$\mathbf{Ae}_j = \lambda_j\mathbf{e}_j \quad \text{and} \quad \mathbf{Ae}_k = \lambda_k\mathbf{e}_k. \qquad (12a,b)$$

Pre-dotting both sides of (12a) with $\mathbf{e}_k$, and post-dotting both sides of (12b) with $\mathbf{e}_j$, and using the fact that $\overline{\lambda} = \lambda$, we have

$$\left.\begin{aligned} \mathbf{e}_k\cdot(\mathbf{Ae}_j) &= \mathbf{e}_k\cdot(\lambda_j\mathbf{e}_j) \\ &= \overline{\lambda}_j(\mathbf{e}_k\cdot\mathbf{e}_j) \\ &= \lambda_j(\mathbf{e}_k\cdot\mathbf{e}_j) \end{aligned}\right| \begin{aligned} (\mathbf{Ae}_k)\cdot\mathbf{e}_j &= (\lambda_k\mathbf{e}_k)\cdot\mathbf{e}_j \\ &= \lambda_k(\mathbf{e}_k\cdot\mathbf{e}_j). \end{aligned} \qquad (13)$$

But (7) tells us that $(\mathbf{A}\mathbf{e}_k) \cdot \mathbf{e}_j = \mathbf{e}_k \cdot (\mathbf{A}^*\mathbf{e}_j)$ and, since $\mathbf{A}^* = \mathbf{A}$ by assumption, the equation on the right side of (13) becomes $\mathbf{e}_k \cdot (\mathbf{A}\mathbf{e}_j) = \lambda_k(\mathbf{e}_k \cdot \mathbf{e}_j)$. Subtracting that equation from the one on the left side of (13) gives

$$0 = (\lambda_j - \lambda_k)(\mathbf{e}_k \cdot \mathbf{e}_j). \tag{14}$$

Finally, $\lambda_j - \lambda_k \neq 0$ since λ_j and λ_k are distinct, by assumption, so it follows from (14) that $\mathbf{e}_k \cdot \mathbf{e}_j = 0$, as was to be shown. ■

THEOREM 12.3.4 *Orthogonal Basis*
If an $n \times n$ matrix $\mathbf{A}$ is Hermitian, then its eigenvectors provide an orthogonal basis for n-space.

Proof of Theorem 12.3.4 follows the same lines as the proof of Theorem 11.3.4.

EXAMPLE 5. Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix}.$$

Setting $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ gives the characteristic equation

$$(3-\lambda)(-\lambda) - (2i)(-2i) = \lambda^2 - 3\lambda - 4 = (\lambda + 1)(\lambda - 4) = 0.$$

Hence,

$$\lambda = -1 \quad \text{and} \quad \lambda = 4.$$

Observe that $\mathbf{A}$ is Hermitian and, sure enough, in accordance with Theorem 12.3.1, the λ's are real.

To find the eigenvectors, solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$.

$$\lambda = \lambda_1 = -1: \quad \begin{bmatrix} 4 & 2i \\ -2i & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{aligned} 4x_1 + 2ix_2 &= 0 \\ -2ix_1 + x_2 &= 0. \end{aligned}$$

Solving, $x_2 = 2ix_1$ so the eigenspace corresponding to λ_1 is

$$\mathbf{e}_1 = \alpha \begin{bmatrix} 1 \\ 2i \end{bmatrix}.$$

Similarly, we find that the eigenspace corresponding to λ_2 is

$$\mathbf{e}_1 = \beta \begin{bmatrix} 2i \\ 1 \end{bmatrix}.$$

Since $\mathbf{A}$ is Hermitian and λ_1 and λ_2 are distinct, the eigenvectors should be orthogonal (Theorem 12.3.3). Let us see: with $\alpha = \beta = 1$, say, $\mathbf{e}_1 \cdot \mathbf{e}_2 = (1)(\overline{2i}) + (2i)(\overline{1}) = -2i + 2i = 0$ so they are, indeed, orthogonal. ■

Hermitian forms, diagonalization, and unitary matrices. The analog of the quadratic form $f(x_1, \dots, x_n) = \mathrm{x}^{\mathrm{T}}\mathbf{A}\mathrm{x}$, where $\mathbf{A}$ is symmetric and $\mathrm{x} = (x_1, \dots, x_n)^{\mathrm{T}}$, is the **Hermitian form**

$$f = \overline{\mathrm{x}}^{\mathrm{T}}\mathbf{A}\mathrm{x}, \tag{15}$$

where $\mathbf{A}$ is Hermitian. For example, if $n = 2$, then (15) becomes

$$\begin{aligned} f &= [\overline{x}_1, \overline{x}_1] \begin{bmatrix} a_{11} & a_{12} \\ \overline{a}_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= a_{11}x_1\overline{x}_1 + a_{22}x_2\overline{x}_2 + a_{12}\overline{x}_1 x_2 + \overline{a}_{12}x_1\overline{x}_2. \end{aligned} \tag{16}$$

The right-hand side of (16) is seen to be real since a_{11} and a_{22} are diagonal elements of an Hermitian matrix and hence are real, $x_1\overline{x}_1 = |x_1|^2$ and $x_2\overline{x}_2 = |x_2|^2$ are real, and the fourth term is the conjugate of the third term. In fact, f is real for *all* values $n \geq 1$:

$$\overline{f} = \overline{\overline{\mathrm{x}}^{\mathrm{T}}\mathbf{A}\mathrm{x}} = \mathrm{x}^{\mathrm{T}}\overline{\mathbf{A}\mathrm{x}} = \overline{(\mathbf{A}\mathrm{x})^{\mathrm{T}}}\mathrm{x} = \overline{\mathrm{x}}^{\mathrm{T}}\overline{\mathbf{A}}^{\mathrm{T}}\mathrm{x} = \overline{\mathrm{x}}^{\mathrm{T}}\mathbf{A}\mathrm{x} = f. \tag{17}$$

(See Exercise 9.)

To reduce f to canonical form, set $\mathrm{x} = \mathbf{U}\tilde{\mathrm{x}}$ in (15). Then

$$f = \overline{(\mathbf{U}\tilde{\mathrm{x}})^{\mathrm{T}}}\mathbf{A}\mathbf{U}\tilde{\mathrm{x}} = \overline{\tilde{\mathrm{x}}}^{\mathrm{T}}(\mathbf{U}^*\mathbf{A}\mathbf{U})\tilde{\mathrm{x}} \tag{18}$$

so $\mathbf{U}$ is to be chosen so that

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} \tag{19}$$

is diagonal. Recalling our discussion of the real case (Section 11.4), it is not surprising that a suitable $\mathbf{U}$ matrix is a normalized modal matrix of $\mathbf{A}$ (Exercise 10). With that choice, the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and f reduces to the canonical form

$$\begin{aligned} f &= \lambda_1\tilde{x}_1\overline{\tilde{x}}_1 + \lambda_2\tilde{x}_2\overline{\tilde{x}}_2 + \cdots + \lambda_n\tilde{x}_n\overline{\tilde{x}}_n \\ &= \lambda_1|\tilde{x}_1|^2 + \lambda_2|\tilde{x}_2|^2 + \cdots + \lambda_n|\tilde{x}_n|^2. \end{aligned} \tag{20}$$

In the real case the normalized modal matrix was denoted as $\mathbf{Q}$, and it turned out that $\mathbf{Q}$ was orthogonal; that is, $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$. Analogously, we find (Exercise 11), in the present case, that

$$\overline{\mathbf{U}}^{\mathrm{T}} = \mathbf{U}^{-1} \quad \text{or} \quad \boxed{\mathbf{U}^* = \mathbf{U}^{-1},} \tag{21}$$

and such a matrix is said to be **unitary**.

THEOREM 12.3.5 *Eigenvalues of Unitary Matrix*
If λ is an eigenvalue of a unitary matrix, then $|\lambda| = 1$.

Proof: Let $\mathbf{U}$ be unitary, with an eigenvalue λ and corresponding eigenvector $\mathbf{e}$:

$$\mathbf{Ue} = \lambda \mathbf{e} \qquad (\mathbf{e} \neq \mathbf{0}). \tag{22}$$

Seeking to employ (21), take the conjugate transpose of both sides:

$$(\overline{\mathbf{Ue}})^{\mathrm{T}} = (\overline{\lambda \mathbf{e}})^{\mathrm{T}} \quad \text{or} \quad \overline{\mathbf{e}}^{\mathrm{T}}\mathbf{U}^{-1} = \overline{\lambda}\,\overline{\mathbf{e}}^{\mathrm{T}}. \tag{23}$$

And post-multiplying the left with $\mathbf{Ue}$ and the right with $\lambda\mathbf{e}$ [which are equal by (22) and nonzero], (23) gives

$$\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{Ue} = \overline{\lambda}\,\overline{\mathbf{e}}^{\mathrm{T}}\lambda\mathbf{e} \quad \text{or} \quad \overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e} = |\lambda|^2\,\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e}. \tag{24}$$

But $\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e} = \|\mathbf{e}\|^2 \neq 0$ so (24) implies that $|\lambda|^2 = 1$, or $|\lambda| = 1$, as claimed. ■

EXAMPLE 6. Reduce

$$f = 3x_1\overline{x}_1 + 2i\overline{x}_1 x_2 - 2ix_1\overline{x}_2 \tag{25}$$

to a canonical form. Comparing (25) with (16), we see that $a_{11} = 3$, $a_{22} = 0$, and $a_{12} = 2i$, so that

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix}.$$

This is the same $\mathbf{A}$ as in Example 5 so

$$\lambda_1 = -1, \ \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ 2i \end{bmatrix}; \qquad \lambda_2 = 4, \ \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{5}}\begin{bmatrix} 2i \\ 1 \end{bmatrix}.$$

Thus, the desired canonical form is

$$f = \lambda_1 |\tilde{x}_1|^2 + \lambda_2 |\tilde{x}_2|^2 = -|\tilde{x}_1|^2 + 4|\tilde{x}_2|^2,$$

where x_1, x_2 and $\tilde{x}_1, \tilde{x}_2$ are related according to

$$\mathbf{x} = \mathbf{U}\tilde{\mathbf{x}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2]\tilde{\mathbf{x}}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2i \\ 2i & 1 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}.$$

Or, if we prefer it the other way around,

$$\tilde{\mathbf{x}} = \mathbf{U}^{-1}\mathbf{x} = \mathbf{U}^{*}\mathbf{x} \quad \text{or} \quad \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & -2i \\ -2i & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \ ■$$

Computer software. Using *Maple*, for instance, no new commands are needed; just type I's for i's. For instance, to find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix},$$

type

```
with(linalg):
```

and return. Then enter

```
A := matrix(2, 2, [3, 2 * I, -2 * I, 0]) :
```

and the command

```
eigenvects(A);
```

The result,

$$[4, 1, \{[2\mathrm{I}, 1]\}], \ [-1, 1, \{[-\frac{1}{2}\mathrm{I}, 1]\}]$$

is the same as found, by hand, in Example 5. The 1's following the eigenvalues 4 and -1 indicate the multiplicity of those eigenvalues.

EXERCISES 12.3

1. Invert each of the following matrices. If the matrix is not invertible (i.e., if it is singular), state that.

(a) $\begin{bmatrix} 2 & i \\ 1 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1+i & 3i \\ 1 & 2-i \end{bmatrix}$

(c) $\begin{bmatrix} 3 & 1+i \\ 1-i & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 3 & i \\ 0 & 1+i & 1 \\ 0 & 0 & 2 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & i & 0 \\ -i & 0 & 2+2i \\ 0 & 2-2i & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 4 & 0 & -i \\ 0 & 4 & 0 \\ i & 0 & 4 \end{bmatrix}$

2. (a)–(f) Same as Exercise 1, but using computer software.

3. (a)–(f) Determine the eigenvalues and eigenvectors of each of the matrices in Exercise 1, and show that the results are in accord with the relevant theorems in this section.

4. (a)–(f) Same as Exercise 3, but using computer software.

5. Determine the eigenvalues and eigenvectors in each case.

(a) $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} 4 & -5 \\ 1 & 2 \end{bmatrix}$

6. Give necessary and sufficient conditions on a, b, c, d such that $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ will have complex eigenvalues, i.e., such that $\operatorname{Im} \lambda \neq 0$.

7. Reduce the following to the canonical form (20), and give the matrix transformations from $\mathbf{x}$ to $\tilde{\mathbf{x}}$, and from $\tilde{\mathbf{x}}$ to $\mathbf{x}$.

(a) $f = 2x_1\bar{x}_1 + 3x_2\bar{x}_2 + (1+i)\bar{x}_1x_2 + (1-i)x_1\bar{x}_2$
(b) $f = 2x_1\bar{x}_1 + 5x_2\bar{x}_2 - 2i\bar{x}_1x_2 + 2ix_1\bar{x}_2$
(c) $f = i\bar{x}_1x_2 - ix_1\bar{x}_2 + 2(1+i)\bar{x}_2x_3 + 2(1-i)x_2\bar{x}_3$
(d) $f = 4x_1\bar{x}_1 + 4x_2\bar{x}_2 + 4x_3\bar{x}_3 + ix_1\bar{x}_3 - i\bar{x}_1x_3$

8. Given each of the following $\mathbf{A}$ matrices, evaluate $\mathbf{A}^{1000}$.

(a) $\begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -2i \\ 2i & 5 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & -i \\ i & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 3 & 2-i \\ 2+i & -1 \end{bmatrix}$

9. Justify the second through fifth equalities in (17), citing relevant equation numbers or theorems.

10. Verify that if $\mathbf{U}$ is a normalized modal matrix of $\mathbf{A}$, then f does reduce to the canonical form (20), as claimed.

11. Verify that a normalized modal matrix $\mathbf{U}$ satisfies (21).

12. Is the Lorentz transformation matrix, in (9), unitary? Orthogonal?

13. If $\mathbf{x} \cdot (\mathbf{Ax}) = \mathbf{x} \cdot (\mathbf{Bx})$ for all $\mathbf{x}$, where $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, does it follow that $\mathbf{A} = \mathbf{B}$? Explain.

14. (*Necessary condition for existence of solutions of* $\mathbf{Ax} = \mathbf{c}$) Prove that a necessary condition for the *existence* of solution(s) to $\mathbf{Ax} = \mathbf{c}$ (where $\mathbf{A}$ is $m \times n$, $\mathbf{x}$ is $n \times 1$, and $\mathbf{c}$ is $m \times 1$) is that $\mathbf{c}$ be orthogonal to every solution $\mathbf{z}$ of the associated homogeneous equation $\mathbf{A}^*\mathbf{z} = \mathbf{0}$. HINT: Dot both sides of $\mathbf{Ax} = \mathbf{c}$ with $\mathbf{z}$, and then use (7).

15. Use the result stated in Exercise 14 to determine necessary conditions, if any, on the components of $\mathbf{c}$, for the given system to be consistent. Compare your result with those obtained by direct application of Gauss elimination to the given system.

(a) $\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 1 \\ 4 & 2 \\ i & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & 3 & 2 & 1 \\ 2 & -1 & 1 & 0 \\ 3 & 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 3 & 2 & i \\ 2 & -1 & 1 & 0 \\ 3 & 2 & 3 & i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(f) $\begin{bmatrix} 2 & 1 & 1 \\ i & 1 & 2i-3 \\ 1 & 2 & -4 \\ -1 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$

16. (*Decomposition*) Given any $m \times n$ matrix $\mathbf{A}$, Hermitian or not, show that $\mathbf{A}$ can be split as $\mathbf{A} = \mathbf{B} + i\mathbf{C}$, where $\mathbf{B}$ and $\mathbf{C}$ are each Hermitian. Show that $\mathbf{B} = (\mathbf{A}^* + \mathbf{A})/2$ and $\mathbf{C} = i(\mathbf{A}^* - \mathbf{A})/2$.

17. (a) If $\mathbf{A}$ is Hermitian, is $i\mathbf{A}$ Hermitian? Explain.
(b) If $\mathbf{A}$ is Hermitian, is $\mathbf{A}^2$ Hermitian? Explain.

Chapter 12 Review

This chapter is so compact that it hardly warrants review. But, let us stress three points.

First, the key difference between $\mathbb{R}^n$ and $\mathbb{C}^n$ is in the *dot product*, which is $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ for $\mathbb{R}^n$ and

$$\boxed{\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{y}}} \tag{1}$$

for $\mathbb{C}^n$, the complex conjugate being introduced so that the dot product $\mathbf{x} \cdot \mathbf{y}$ and the norm $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$ satisfy the properties listed in equations (12) and (17), respectively, in Section 9.5. (If you studied the optional Section 9.6 you will recall that these properties were elevated to requirements, or axioms, for any normed inner product space.)

Second, just as real matrices are found, in Section 11.3, to admit special useful properties (e.g., the eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and the eigenvectors provide an orthogonal basis for the n-space) if they are symmetric ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$), we find in this chapter that complex matrices admit the same properties if they are *Hermitian* ($\mathbf{A}^* = \mathbf{A}$, where $\mathbf{A}^*$ is $\overline{\mathbf{A}}^{\mathrm{T}}$).

Finally, we call your attention to the result

$$\boxed{(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y}),} \tag{2}$$

given as equation (7) in Section 12.3. For real matrices the latter becomes

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}\mathbf{y}) \tag{3}$$

if $\mathbf{A}$ is symmetric, and for complex matrices (3) holds if $\mathbf{A}$ is Hermitian; (3) is the key relation used in proving that eigenvectors corresponding to distinct eigenvalues are orthogonal, which result, in turn, is needed in proving that the eigenvectors provide an orthogonal basis for the n-space. We will meet a function-space version of (3) known as *Lagrange's identity*, in Chapter 17, when we study the *Sturm–Liouville theory*. Expansions in terms of bases consisting of orthogonal eigenvectors (or "eigenfunctions" in the function space case) will be of great importance to us, and therefore the underlying relation (3) is of great importance as well.

Chapter 13

Differential Calculus of Functions of Several Variables

13.1 Introduction

In Chapters 8 – 12 on Linear Algebra most of our interest was in n-space. In Chapters 13–16 on Multivariable Calculus and Field Theory we return to physical two- and three-dimensional space. The heart of this group of chapters is Chapter 16, on scalar and vector field theory. The three preceding chapters prepare the way by covering a number of topics from the subject area generally known as "advanced calculus."

In Chapter 13 we consider (real valued) functions of more than one (real) variable, and lay some of the groundwork for our subsequent development of field theory and partial differential equations. Much of the development parallels that in the calculus of functions of a single variable: derivatives, chain differentiation, Taylor's formula, the mean value theorem, and maxima and minima.

Chapter 14 is not a continuation of the development begun in Chapter 13; it covers the fundamentals of vectors and their manipulation: the dot and cross product, base vectors, differentiation of vectors, Cartesian and non-Cartesian coordinate systems, and the representation of curves in space. Chapter 15 returns to the calculus of functions of several variables, begun in Chapter 13, with introductory material on double and triple integrals, and on surfaces and volumes. Finally, in Chapter 16 we consider both scalar and vector fields; we introduce the divergence, gradient, and curl, and consider several integral theorems and a number of physical applications from the subjects of heat conduction, fluid flow, and electrostatics.

Discussion of non-Cartesian coordinates (polar, cylindrical, and spherical) is confined to OPTIONAL sections so that it can be omitted in a shorter course.

13.2 Preliminaries

13.2.1. Functions. We will be considering real-valued functions of n real variables, say $f(x_1, \ldots, x_n)$. The function is a **mapping**, or **transformation**, from a point $\mathbf{x} = (x_1, \ldots, x_n)$ on an f axis, $\mathbb{R}^1$, as illustrated in Fig. 1a for $n = 1$. To obtain a graphical display of the mapping one can draw arrows from a number of x points to their image points on the f axis as we have done in Fig. 1b for the function $f(x) = x^2$. However, Descartes had a better idea: place the x and f axes at right angles to each other and plot all points $(x, f(x))$ for the desired x interval. That procedure gives the familiar **graph** of f, as in Fig. 1c.

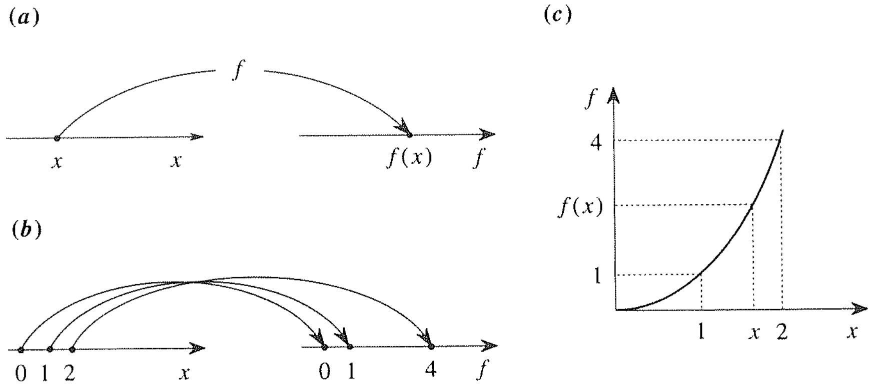


Figure 1. The function f.

Similarly, if $n = 2$ we can obtain the graph of f as a surface in three-dimensional Cartesian x_1, x_2, f space. Of course, if $n \geq 3$ the idea fails because we would need four or more dimensions in which to present the graph.

In Chapters 13–16, n will usually be 2 or 3, but there is no need for any such limitation in the present discussion.

13.2.2. Point set theory definitions. The physical nature of the independent variables $x_1, \ldots, x_n$ is irrelevant here; they may be pressures, distances, or whatever. As noted above, we will regard $x_1, \ldots, x_n$ as the coordinates of a point in an n-dimensional space and will denote that point as $\mathbf{x} = (x_1, \ldots, x_n)$. Further, we define the **distance** $d(\mathbf{x}, \mathbf{x}')$ between $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{x}' = (x_1', \ldots, x_n')$ as

$$\boxed{d(\mathbf{x}, \mathbf{x}') \equiv \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2}.} \tag{1}$$

If $n = 1$, then $\mathbf{x}$ and $\mathbf{x}'$ are two points x_1 and x_1' on an x axis, and (1) becomes

$$d(\mathbf{x}, \mathbf{x}') = |x_1 - x_1'|\,. \tag{2a}$$

For $n = 2$

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}, \tag{2b}$$

and for $n = 3$

$$d(\mathrm{x}, \mathrm{x}') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_2 - x_3')^2}. \qquad \text{(2c)}$$

These formulas are familiar to us: (2b) and (2c) are the Pythagorean formulas for the distance between points in two- and three-dimensional Euclidean space, and for $n > 3$ the definition (1) amounts to an n-dimensional generalization of these Pythagorean formulas.*

Next, we define a **neighborhood** $N(\mathrm{x}'; r)$ of a point x' as the set of all points x closer to x' than r, namely, such that

$$d(\mathrm{x}, \mathrm{x}') < r. \qquad (r > 0) \qquad \text{(3)}$$

In one dimension (3) becomes $|x' - x| < r$, which corresponds to the *interval* shown in Fig. 2a (not including the endpoints $x' + r$ and $x' - r$), in two dimensions N is the circular *disk* shown in Fig. 2b (not including the points on its edge), in three dimensions N is a *sphere* (not including the points on its surface), and if $n > 3$ we speak of N as an n-dimensional *hypersphere*.

Next, we say that a point set S is **connected** if each pair of points in S can be joined by an unbroken line consisting of a finite number of straight segments, each contained entirely within S.†

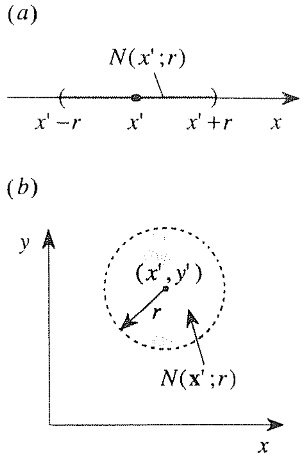


Figure 2. One- and two-dimensional neighborhoods.

EXAMPLE 1. Each of the point sets S_1 and S_2 in the two-dimensional plane, shown in Fig. 3, is connected. The points $\mathrm{x}_1, \mathrm{x}_2$, for example, may be joined as shown. In contrast,

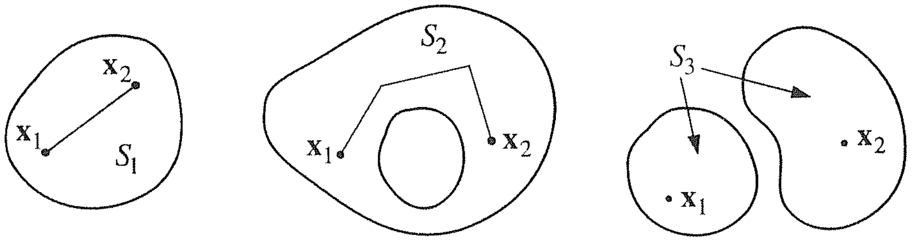


Figure 3. Connectedness.

*Recall from Section 9.5 that (1) is the Euclidean norm of the vector from x to x' in n-space. It is not the *only* definition of distance possible (just as the Euclidean norm is not the only viable norm in n-space), but it is the one most commonly adopted. For instance, $d(\mathrm{x}, \mathrm{x}') \equiv |x_1 - x_1'| + \cdots + |x_n - x_n'|$ is also acceptable and is sometimes preferred because it is algebraically simpler.

†A straight line may be defined in n-dimensional space, even if $n > 3$, by parametric equations $x_1 = a_1 + b_1 t, x_2 = a_2 + b_2 t, \ldots, x_n = a_n + b_n t$, where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are constants, and where t is the parameter.

the set S_3 is not connected since not *all* pairs of points in S_3 can be joined by linear segments lying entirely within S_3. The points $\mathbf{x}_1, \mathbf{x}_2$, for example, cannot be so joined. ▮

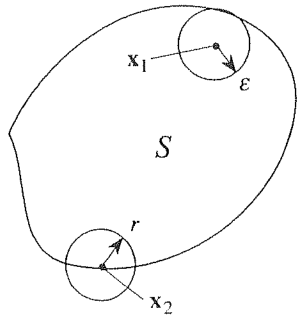


Figure 4. Interior point $\mathbf{x}_1$, boundary point $\mathbf{x}_2$.

A point $\mathbf{x}$ is said to be a **boundary point** of a point set S if *every* neighborhood $N(\mathbf{x}; r)$ (i.e., no matter how small we take r to be) contains points in S *as well as* points not in S. Thus, a boundary point is just what it sounds like. To illustrate, the point $\mathbf{x}_1$ in the two-dimensional point set S shown in Fig. 4 is *not* a boundary point since the disk $N(\mathbf{x}_1; r)$ contains no points outside S if we choose r to be smaller than ϵ. The point $\mathbf{x}_2$, on the other hand, *is* a boundary point since $N(\mathbf{x}_2; r)$ evidently contains points in S *and* points not in S, no matter how small we take r to be. (Recall, from the definition of neighborhood, that $r > 0$, so r cannot be zero.)

A point $\mathbf{x}$ of a set S is said to be an **interior point** of S if there exists some neighborhood of $\mathbf{x}$ lying entirely within S. For instance, the point $\mathbf{x}_1$ in Fig. 4 is an interior point of S because the disk $N(P; r)$ lies entirely within S for any $r < \epsilon$.

Finally, we call a connected set S an **open region** or **domain** if it contains *none* of its boundary points, and a **closed region** if it contains *all* of its boundary points.

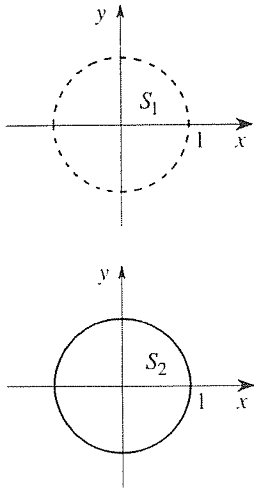


Figure 5. Open and closed regions.

EXAMPLE 2. Let S_1 be the set of points (x, y), in Cartesian 2-space, such that $x^2 + y^2 < 1$ as shown in Fig. 5. We write $S_1 = \{(x, y) | \, x^2 + y^2 < 1\}$. Evidently, S_1 is connected. Its boundary points are the points on the unit circle $x^2 + y^2 = 1$. None of these points belongs to S_1 (as conveyed in the figure by the use of dashed lines), so S_1 is an open domain. We call this domain the *open unit disk*. Every point in S_1 is an interior point.

Next, let $S_2 = \{(x, y) | \, x^2 + y^2 \leq 1\}$ as shown in Fig. 5. This time all the boundary points (those on the circle $x^2 + y^2 = 1$) are contained in S_2, as emphasized by the solid boundary line in the figure, so S_2 is a closed region, the *closed unit disk*. All points (x, y) such that $x^2 + y^2 < 1$ are interior points of S_2. ▮

Of course, if S contains some but not all of its boundary points, then it is neither an open region *nor* a closed region. However, the regions that we shall work with will generally be either open or closed. Further, the terms "open" and "closed" are not mutually exclusive. For example, the connected set $S = \{x | -\infty < x < \infty\}$, namely, the entire real axis, has no boundary points. Thus it is at once open (since it contains none of its boundary points) and closed (since it contains all of its boundary points).

13.2.3. Limits and continuity. To define limits and continuity for functions of n variables, let us first review these concepts, from the calculus, for a function of a single variable.

We say that $f(x)$ has a **limit** L as x tends to x', and write

$$\lim_{x \to x'} f(x) = L \quad \text{or} \quad f(x) \to L \quad \text{as } x \to x', \tag{4}$$

if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, x') > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x'| < \delta$. That is, $f(x)$ can be made arbitrarily close to L by making x sufficiently close to x'.

EXAMPLE 3. Consider $f(x) = 1/x$ over $0.1 \le x \le 1$. At the x' shown in Fig. 6, the limit exists and is equal to the L shown, namely, $1/x'$. To prove that claim we need to put forward a suitable function $\delta(\epsilon, x')$. Draw an arbitrarily small ϵ band about L and, where it intersects the graph (at A and B), drop verticals to the x axis. Observing that $a < b$, we can choose $\delta = a$. Then $|x - x'| < \delta$ is the centered interval denoted by the small parentheses. Surely, if x is closer to x' than δ (i.e., $0 < |x - x'| < \delta$), then $f(x)$ will be closer to L than ϵ, as desired. To determine the functional form of δ, write $f(x' - \delta) - f(x') = \epsilon$. That is, $1/(x' - \delta) - 1/x' = \epsilon$. Solving for δ,

$$\delta(\epsilon, x') = x' - \frac{x'}{\epsilon x' + 1}. \tag{5}$$

Of course any smaller (nonzero) value, such as $\frac{1}{5}[x' - x'/(\epsilon x' + 1)]$, will do just as well.

COMMENT. Practically speaking, we don't really need to put forward a $\delta(\epsilon, x')$ to convince ourselves that the limit of $f(x)$, as $x \to x'$, is L (in Fig. 6). Our purpose, in doing so here, has been to clarify the meaning of the ϵ, δ definition of limit, given above. ▮

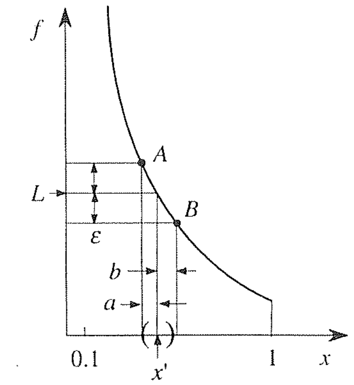


Figure 6. $\lim f(x)$ for $f(x) = 1/x$.

Note carefully that because of the "$0 <$" in $0 < |x - x'| < \delta$, in our definition of $\lim_{x \to x'} f(x)$, $f(x')$ need not equal L. For instance, if

$$f(x) = \begin{cases} x^2, & 0 \le x \le 3 \quad \text{but } x \ne 2 \\ 12, & x = 2 \end{cases} \tag{6}$$

then (Fig. 7) $\lim_{x \to 2} f(x) = 4$, not 12. [Similarly, the limit of $f(x)$ is 0 as $x \to 0$, 1 as $x \to 1$, 6.25 as $x \to 2.5$, and 9 as $x \to 3$.][*]

If, *in addition* to having $\lim_{x \to x'} f(x) = L$ we also have $f(x') = L$, hence

$$\lim_{x \to x'} f(x) = f(x'), \tag{7}$$

then we say that $f(x)$ is **continuous** at x'. In ϵ, δ language, *$f(x)$ is continuous at x' if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, x') > 0$ such that $|f(x) - f(x')| < \epsilon$ whenever $|x - x'| < \delta$.* If $f(x)$ is not continuous, then it is **discontinuous**.

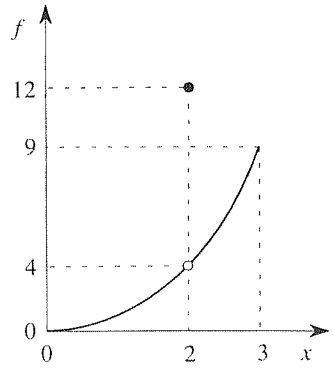


Figure 7. The function defined by (6).

For instance, the function $f(x) = 1/x$ is continuous on any x interval not containing the origin, such as $1 \le x < \infty$, $0 < x \le 5$, or $-\infty < x < 0$. The functions e^x and $\sin x$ are continuous on $-\infty < x < \infty$, and the function defined

[*] In $0 < |x - x'| < \delta$, within the italicized definition of limit, it is understood that x also needs to be within the interval of definition of the function. Thus, we say that $f(x) \to 6.25$ as $x \to 2.5$ is a *two-sided limit*, since x can approach 2.5 from the right or from the left, but $f(x) \to 9$ as $x \to 3$ is a *one-sided limit*, or a *left-hand limit*, since x can approach 3 only from the left. Similarly, $f(x) \to 0$ as $x \to 0$ is a one-sided or *right-hand limit*.

by (6) and shown in Fig. 7 is continuous on $0 \leq x \leq 3$, except at $x = 2$. The Heaviside step function, defined by

$$H(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases} \tag{8}$$

with $H(0) = \frac{1}{2}$, say, is continuous on any interval not containing the origin. At the origin it is discontinuous because (7) does not hold there: $f(x') = f(0) = \frac{1}{2}$, but the left-hand side of (7) does not exist because $f(x) \to 1$ as $x \to 0$ from the right, whereas $f(x) \to 0$ as $x \to 0$ from the left. For $\lim_{x \to x'} f(x)$ to exist, the limit from the left (the *left-hand limit*) and the limit from the right (the *right-hand limit*) must exist and equal each other. In the present example both of those limits exist, but they do not equal each other.

Similarly for the functions of several variables. We say that $f(\mathbf{x})$ is continuous at x' if

$$\lim_{\mathbf{x} \to \mathbf{x}'} f(\mathbf{x}) = f(\mathbf{x}'), \tag{9}$$

where $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{x}' = (x_1', \ldots, x_n')$, and where we use $f(\mathbf{x})$ for brevity to denote $f(x_1, \ldots, x_n)$, and similarly for $f(\mathbf{x}')$. In ϵ, δ language, $f(\mathbf{x})$ is continuous at $\mathbf{x}'$ if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, \mathbf{x}')$ such that $|f(\mathbf{x}) - f(\mathbf{x}')| < \epsilon$ whenever $d(\mathbf{x}, \mathbf{x}') < \delta$, where $d(\mathbf{x}, \mathbf{x}')$ was defined by (1). If it is not continuous, then it is discontinuous.

For instance, $f(x_1, x_2) = x_1^2 + 3 \sin(x_1 x_2)$ is continuous for all values (i.e., for all finite values) of x_1 and x_2, and $g(x_1, x_2, x_3) = 1/(x_1^2 + x_2^2 + x_3^2)$ is continuous everywhere except at $x_1 = x_2 = x_3 = 0$. Generally, the functions that arise in applications are either continuous everywhere, or almost everywhere. Of the example functions f and g, for instance, f is continuous everywhere and g is continuous everywhere except at the origin.

Closure. To study limits and continuity one first needs to establish basic concepts of point set theory. For instance, if we wish $\lim_{\mathbf{x} \to \mathbf{x}'} f(\mathbf{x}) = L$ to mean that $f(\mathbf{x})$ can be made arbitrarily close to L by making $\mathbf{x}$ sufficiently close to $\mathbf{x}'$, then we need to make the concept of closeness precise by defining the *distance* between $\mathbf{x}$ and $\mathbf{x}'$. We adopt the n-dimensional version

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{(x_1 - x_1')^2 + \cdots + (x_n - x_n')^2}$$

of the familiar two- and three-dimensional Euclidean distance. Other terms defined here, that will come up in these chapters, include neighborhood, connected, boundary point, interior point, open region (or domain), and closed region.

EXERCISES 13.2

1 Determine the distance $d(P, P')$ in each case.

(a) $P = (3, 1, 0)$, $P' = (-1, 2, 2)$
(b) $P = (1, -1, 5, 0)$, $P' = (0, 4, 3, 2)$
(c) $P = (1, -4)$, $P' = (6, 5)$
(d) $P = (0, 4, 3, 0)$, $P' = (0, 0, 0, 5)$
(e) $P = (-6)$, $P' = (-2)$
(f) $P = (8, 0, 1, 0)$, $P' = (0, 2, -6, 7)$

2. Is $P = (3.2, 3.7)$ in $N(P_0; r)$ if $P_0 = (3, 4)$ and $r = 0.4$?

3. Is $P = (4.3, 1.1)$ in $N(P_0; r)$ if $P_0 = (4.2, 1)$ and $r = 0.2$?

4. Is $P = (2, 5, 7)$ in $N(P_0; r)$ if $P_0 = (3, 4, 5)$ and $r = 3$?

5. Is $P = (0.99, 1.96, -0.95, 0.03)$ in $N(P_0; r)$ if $P_0 = (1, 2, -1, 0)$ and $r = 0.01$?

6. In each case identify the boundary points and indicate whether or not the set is connected, an open region, a closed region, or neither.

(a) $\{x \mid 0 < x \leq 1\}$
(b) $\{x \mid -3 \leq x < 5\}$
(c) $\{x \mid 0 < x < \infty\}$
(d) $\{x \mid 0 \leq x < \infty\}$
(e) $\{x \mid 0 < \sin x \leq 1\}$
(f) $\{x \mid \sin x = 0\}$
(g) $\{(x_1, x_2) \mid 2 < x_1 < 3, 1 < x_2 < 5\}$
(h) $\{(x_1, x_2) \mid -\infty < x_1 < \infty, 0 < x_2 < 1\}$
(i) $\{(x_1, x_2) \mid 6 < x_1 < 8, x_2 = 0\}$
(j) $\{(x_1, x_2) \mid 1 < x_1^2 + x_2^2 \leq 4\}$
(k) $\{(x_1, x_2) \mid 0 < x_1^2 + x_2^2 < 1\}$
(l) $\{(x_1, x_2) \mid -2 < x_1 x_2 < 2\}$
(m) $\{(x_1, x_2, x_3) \mid 1 < x_1^2 + x_3^2 < 2, x_2 = 0\}$
(n) $\{(x_1, x_2, x_3, x_4) \mid 3 \leq x_1^2 + x_2^2 + x_3^2 + x_4^2 < \infty\}$
(o) $\{(x_1, x_2, x_3, x_4) \mid 0 < x_1 < 1, 0 < x_2 < 1, 0 < x_3 < 1, 0 < x_4 < 1\}$

7. For the stated limit, put forward a suitable $\delta(\epsilon)$, as we did in Example 3. NOTE: In Example 3, x' was any point in $0.1 \leq x \leq 1$, so δ depended not only on ϵ but also on x'. In this exercise x' is a specific point, so δ will depend only on ϵ.

(a) $\lim_{x\to 2} x^2 = 4$
(b) $\lim_{x\to 5} 3x = 15$
(c) $\lim_{x\to 1} \sin x = \sin 1$
(d) $\lim_{x\to 2} \sin x = \sin 2$
(e) $\lim_{x\to 0} e^x = 1$
(f) $\lim_{x\to 0} e^{-x} = 1$

8. (a) Does $\lim_{x\to 0} \sin(1/x)$ exist? If so, give its value; if not, explain why not. HINT: To sketch the graph of $\sin(1/x)$ it is useful to re-express it as $\sin[(1/x^2)x]$ since then we have the more familiar form $\sin \omega x$, where in this case the frequency $\omega = 1/x^2$ is a function of x. Alternatively, setting $t = 1/x$ observe that $\lim_{x\to 0} \sin(1/x) = \lim_{t\to\infty} \sin t$.
(b) Show that $\lim_{x\to 0} x^2 \sin(1/x) = 0$, by putting forward a suitable $\delta(\epsilon)$.

9. (a) Prove that

$$|A + B| \leq |A| + |B|. \tag{9.1}$$

(b) Show that it follows from (9.1) that

$$|A_1 + \cdots + A_n| \leq |A_1| + \cdots + |A_n|. \tag{9.2}$$

(c) Prove that $\lim_{x\to a} f(x) = A$ is equivalent to the statement

$$\lim_{x\to a} [f(x) - A] = 0. \tag{9.3}$$

(d) Prove that

$$\lim_{x\to a} [f(x) + g(x)] = \lim_{x\to a} f(x) + \lim_{x\to a} g(x), \tag{9.4}$$

provided that each of the limits on the right exists. HINT: Let $\lim_{x\to a} f(x) = A$ and $\lim_{x\to a} g(x) = B$. Then, for any $\epsilon > 0$ there is a δ_1 such that $|f(x) - A| < \epsilon$ whenever $0 < |x - a| < \delta_1$, and a δ_2 such that $|g(x) - B| < \epsilon$ whenever $0 < |x - a| < \delta_2$. Then, express $|[f(x) + g(x)] - (A + B)|$ as $|[f(x) - A] + [g(x) - B]|$ and use (9.1).
(e) Prove that

$$\lim_{x\to a} [Cf(x)] = C \lim_{x\to a} f(x), \tag{9.5}$$

provided that $\lim_{x\to a} f(x)$ exists.
(f) Show that it follows from (9.4) and (9.5) that the limit is **linear**:

$$\lim_{x\to a} [\alpha f(x) + \beta g(x)] = \alpha \lim_{x\to a} f(x) + \beta \lim_{x\to a} g(x) \tag{9.6}$$

for arbitrary constants α, β, provided that each of the limits on the right exists.
(g) Prove that

$$\lim_{x\to a} [f(x)g(x)] = [\lim_{x\to a} f(x)][\lim_{x\to a} g(x)], \tag{9.7}$$

provided that each of the limits on the right exists. HINT: Let $\lim_{x\to a} f(x) = A$ and $\lim_{x\to b} g(x) = B$. Then, for any $\epsilon > 0$ there is a δ_1 such that $|f(x) - A| < \epsilon$ whenever $0 < |x - a| < \delta_1$, and there is a δ_2 such that $|g(x) - B| < \epsilon$

whenever $0 < |x - a| < \delta_2$. Let ϵ be less than unity. Then, $|f(x)| = |A + [f(x) - A]| < |A| + \epsilon < |A| + 1$ so

$$\begin{aligned} |f(x)g(x) - AB| &= |f(x)g(x) - Bf(x) + Bf(x) - AB| \\ &\leq |f(x)||g(x) - B| + |B||f(x) - A| \\ &< (|A| + 1)\epsilon + |B|\epsilon. \end{aligned} \tag{9.8}$$

Provide the remaining steps in the proof.

10. We state in Exercise 9 that (9.4) holds, "provided that each of the limits on the right exists." Give a counterexample, where (9.4) fails to hold because the limit on the left exists, yet those on the right do not. That is, give $f(x)$, $g(x)$, and a.

11. We stated that (9.5) in Exercise 9 holds, "provided that $\lim_{x\to a} f(x)$ exists." Give a counterexample, where (9.5) fails to hold because the limit on the left exists, yet the one on the right does not. That is, give $f(x)$, C, and a.

12. Identify any points or point sets (for instance, the plane $x + y - 2z = 5$), if any, where f is discontinuous.

(a) $f(x) = \dfrac{x+1}{x(x^2 + x - 2)^2}$
(b) $f(x) = \tan x$
(c) $f(x, y) = \tan(x + y)$
(d) $f(x, y) = e^x/(y + 3)$
(e) $f(x, y) = (x + y)/(x^2 + y)^2$
(f) $f(x, y, z) = x/(yz)$
(g) $f(x, y, z) = 5/(x^2 + y^2 + z^2 - 1)$

13.3 Partial Derivatives

To review partial differentiation, it should suffice to consider a function f of two variables, say x and y. Suppose $f(x, y)$ is defined throughout some neighborhood of a point $\mathbf{x}_0 = (x_0, y_0)$. Let us hold y fixed at y_0, so that $f(x, y) = f(x, y_0)$ is then a function of x alone. If the x-derivative of this function exists at x_0, it is called the **partial derivative** of f with respect to x, at (x_0, y_0), and is usually denoted either as $\partial f/\partial x$ or as f_x. Thus,

$$f_x = \frac{\partial f}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x} \tag{1a}$$

at (x_0, y_0). Similarly,

$$f_y = \frac{\partial f}{\partial y} = \lim_{\Delta y \to 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y} \tag{1b}$$

at (x_0, y_0). For example, if $f(x, y) = x^2 y^3$, then $f_x = 2xy^3$ and $f_y = 3x^2 y^2$ at any given point (x, y).

In geometrical terms, (1a) is the slope of the tangent line A, and (1b) is the slope of the tangent line B in Fig. 1. For the case shown in the figure, $f_x > 0$ and $f_y < 0$ at $\mathbf{x}_0$, that is, at (x_0, y_0).

If the partial derivatives f_x and f_y exist not only at $\mathbf{x}_0$, but throughout some neighborhood of $\mathbf{x}_0$, then they are functions of x and y which may, in turn, admit further derivatives, namely, the *second-order partial derivatives*

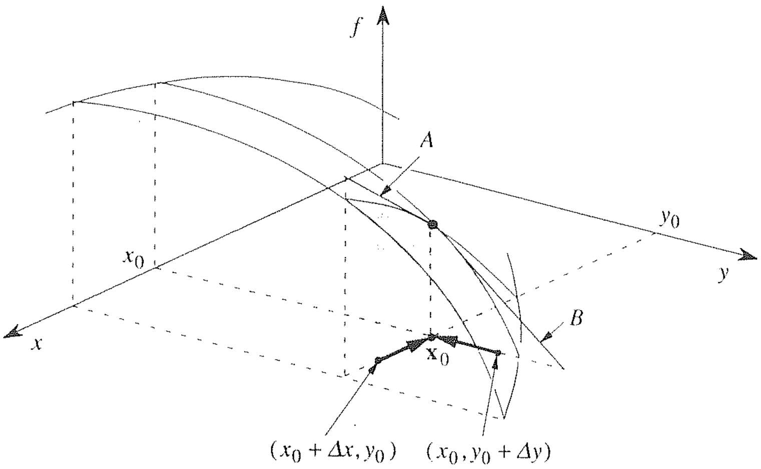


Figure 1. f_x and f_y at $\mathbf{x}_0$.

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) &= \frac{\partial^2 f}{\partial x^2} = f_{xx}, & \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) &= \frac{\partial^2 f}{\partial x \partial y} = f_{yx}, \\ \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) &= \frac{\partial^2 f}{\partial y \partial x} = f_{xy}, & \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) &= \frac{\partial^2 f}{\partial y^2} = f_{yy}. \end{aligned} \tag{2}$$

Similarly, f may admit *third-order partial derivatives*, such as

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial y^2}\right) = \frac{\partial^3 f}{\partial x \partial y^2} = f_{yyx},$$

and so on.

Note the order of the subscripts. For example, f_{yx} means $(f_y)_x$. Thus,

$$f_{yx} = (f_y)_x = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \qquad f_{xyy} = ((f_x)_y)_y = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)\right),$$

and so on. Does the order really matter? That is, for a "mixed" derivative such as f_{xy} is it true that $f_{xy} = f_{yx}$ always? Consider two examples.

EXAMPLE 1. Let $f(x, y) = x \sin(xy^2)$. Then

$$\begin{aligned} f_x &= \sin(xy^2) + xy^2 \cos(xy^2), \\ f_{xy} &= 2xy \cos(xy^2) + 2xy \cos(xy^2) - 2x^2y^3 \sin(xy^2), \\ f_y &= 2x^2 y \cos(xy^2), \\ f_{yx} &= 4xy \cos(xy^2) - 2x^2 y^3 \sin(xy^2), \end{aligned}$$

so that $f_{xy} = f_{yx}$, in this case, for all values of x and y. ▮

EXAMPLE 2. Let

$$f(x,y)=\begin{cases}\dfrac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y)\neq(0,0)\\ 0 & \text{if } (x,y)=(0,0).\end{cases} \tag{3a,b}$$

[The separate specification (3b) is needed because at the origin (3a) gives $0/0$, which is not defined; (3b) was *not* obtained from (3a) by letting x and y tend to zero, it was stated as the author's choice, as part of the *definition* of $f(x,y)$.] If we stay away from the origin, and work only with (3a), we would find, as in Example 1, that $f_{xy}=f_{yx}$ at each point (x,y) in the plane. We will demonstrate that $f_{xy}=f_{yx}$ does *not* hold, however, at the origin $(0,0)$.

First, take $\partial/\partial x$ of f. That step gives

$$f_x(x,y)=\begin{cases}\dfrac{y(x^4+4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y)\neq(0,0)\\ 0 & \text{if } (x,y)=(0,0).\end{cases} \tag{4a,b}$$

Obtaining (4a) from (3a) was straightforward, but where did we get (4b)? We obtained (4b) directly from (1a), as follows:

$$\begin{aligned}f_x(0,0)&=\lim_{\Delta x\to 0}\frac{f(\Delta x,0)-f(0,0)}{\Delta x}\\&=\lim_{\Delta x\to 0}\frac{0-0}{\Delta x}=\lim_{\Delta x\to 0}0=0,\end{aligned}$$

where $f(\Delta x,0)=0$ followed from (3a) and $f(0,0)=0$ from (3b). Continuing,

$$\begin{aligned}f_{xy}(0,0)&=\lim_{\Delta y\to 0}\frac{f_x(0,\Delta y)-f_x(0,0)}{\Delta y}\\&=\lim_{\Delta y\to 0}\frac{-\Delta y-0}{\Delta y}\quad\text{from (4a) and (4b)}\\&=-1.\end{aligned} \tag{5}$$

Similarly, we find that

$$f_y(x,y)=\begin{cases}\dfrac{x(x^4-4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y)\neq(0,0)\\ 0 & \text{if } (x,y)=(0,0)\end{cases} \tag{6a,b}$$

and

$$\begin{aligned}f_{yx}(0,0)&=\lim_{\Delta x\to 0}\frac{f_y(\Delta x,0)-f_y(0,0)}{\Delta x}\\&=\lim_{\Delta x\to 0}\frac{\Delta x-0}{\Delta x}\quad\text{from (6a) and (6b)}\\&=1.\end{aligned} \tag{7}$$

Comparing (5) and (7), we see that $f_{xy}(0,0)\neq f_{yx}(0,0)$ in this example. ∎

In general, then, the order of differentiation does matter. The following theorem gives conditions which guarantee that the mixed partial derivatives f_{xy} and

f_{yx} of a given function $f(x, y)$ are equal.

THEOREM 13.3.1 *Equality of Mixed Partials*
If f_x, f_y, f_{xy}, and f_{yx} are continuous in some neighborhood of (x_0, y_0), then $f_{yx} = f_{xy}$ at (x_0, y_0).*

Evidently, the function $f(x, y)$ in Example 2 must not have met all of the conditions of Theorem 13.3.1; see Exercise 6.

Closure. Theorems analogous to Theorem 13.3.1 can be obtained for mixed partial derivatives of higher order as well, but we will not go into the details here because the vast majority of functions that are met in applications are sufficiently well behaved to ensure that the order of differentiation is immaterial. In that case, why discuss the matter at all? To bring this important theoretical question to light and, at the very least, to answer it for the important case of mixed partial derivatives of second order.

Further, Theorem 13.3.1 is representative of numerous theorems that deal with whether or not it is permissible to interchange the order of two *limit* operations. For instance, is

$$\iint_{\mathcal{D}} f(x, y)\, dxdy = \iint_{\mathcal{D}} f(x, y)\, dydx\,? \tag{8}$$

Is

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}\,? \tag{9}$$

Is

$$\frac{d}{dt}\int_a^b f(x, t)\, dx = \int_a^b \frac{\partial f}{\partial t}(x, t)\, dx\,? \tag{10}$$

That is, does it matter if we integrate first on x and then on y or vice versa? Or if we sum first on j and then on i or vice versa? Or if we first integrate with respect to x and then differentiate the result with respect to t or vice versa? Each of these operations is indeed a limit operation: integration involves the limit of a sequence of Riemann sums, an infinite sum involves the limit of a sequence of partial sums, and differentiation involves the limit of a difference quotient.

The general idea is that the interchange can be performed with impunity if the quantity involved [$f(x, y)$ in (8), a_{ij} in (9), $f(x, t)$ in (10)] is sufficiently well behaved. For instance, (10) holds if $f(x, t)$ and $\partial f(x, t)/\partial t$ are continuous over the

*For proof, see J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Francisco: W. H. Freeman, 1976), pg. 120. They ask a bit more in their theorem, namely, that f_{xx} and f_{yy} be continuous too. But they do not use those conditions in their proof, which therefore holds for our theorem as well.

region of interest in the x, t plane. The relevant theorems can be found in textbooks on advanced calculus.*

EXERCISES 13.3

1. Evaluate f_x, f_y, f_{xy}, f_{yx}, f_{xx}, and f_{yy}. Are there any points (x, y) at which $f_{xy} \neq f_{yx}$? Explain.

(a) $f = x^3y^5$ (b) $f = y \sin(3x^2y)$
(c) $f = 1/(x^2 + y^2 + 1)$ (d) $f = x^3 + y^5$
(e) $f = (x^2 + y^2)^{1/2}$ (f) $f = (x^2 + y^2)^{3/2}$

2. Verify that $f_{xxy} = f_{xyx} = f_{yxx}$ everywhere in the x, y plane.

(a) $f = x^4y^3$ (b) $f = \cos(x^2y)$
(c) $f = \exp(x^2 + y^2)$ (d) $f = \sin(x - y^2)$

3. Recall that PDE is shorthand for partial differential equation. Show that

(a) $f = \ln(x^2 + y^2)$ satisfies the PDE $f_{xx} + f_{yy} = 0$, except at $(0, 0)$
(b) $f = \ln[(x - x_0)^2 + (y - y_0)^2]$ satisfies the PDE $f_{xx} + f_{yy} = 0$, except at (x_0, y_0)
(c) $f = r^n \sin n\theta$ satisfies the PDE $r^2 f_{rr} + rf_r + f_{\theta\theta} = 0$ for $n = 0, \pm 1, \pm 2, \ldots$ everywhere, except at $r = 0$ for the case where n is negative
(d) $f = r^n \cos n\theta$ satisfies the PDE $r^2 f_{rr} + rf_r + f_{\theta\theta} = 0$ for $n = 0, \pm 1, \pm 2, \ldots$ everywhere, except at $r = 0$ for the case where n is negative
(e) $f = 1/[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]$ satisfies the PDE $f_{xx} + f_{yy} + f_{zz} = 0$, except at (x_0, y_0, z_0)
(f) $f = \sin \kappa x \exp(-\kappa^2 t)$ satisfies the PDE $f_{xx} = f_t$ for all (x, t), where κ is any constant
(g) $f = \cos \kappa x \exp(-\kappa^2 t)$ satisfies the PDE $f_{xx} = f_t$ for all (x, t), where κ is any constant
(h) $f = \sin(x - ct)$ satisfies the PDE $c^2 f_{xx} = f_{tt}$ for all (x, t), where c is any constant

4. Determine the allowable values, if any, of the constant α such that $f = (x_1^2 + \cdots + x_n^2)^\alpha$ is a solution of the PDE

$$f_{x_1x_1} + \cdots + f_{x_nx_n} = 0 \tag{4.1}$$

[everywhere in $(x_1, \cdots, x_n)$ space, except possibly at the origin].

5. Let $f(x, y) = x$ for $y = 0$ and 0 for $y \neq 0$. Evaluate each of the following; if they do not exist, state that.

(a) $f_x(0, 0)$ (b) $f_x(0, 2)$ (c) $f_x(2, 0)$
(d) $f_x(3, 4)$ (e) $f_y(6, 0)$ (f) $f_y(0, 2)$
(g) $f_y(0, 0)$ (h) $f_{yx}(0, 0)$ (i) $f_{xy}(0, 0)$
(j) $f_{xy}(3, 0)$ (k) $f_{yxx}(3, 0)$ (l) $f_{xxy}(3, 0)$

6. In Example 2 we found that $f_{xy}(0, 0) \neq f_{yx}(0, 0)$. Evidently, the function $f(x, y)$ defined by (3) does not meet all the requirements stated in Theorem 13.3.1.

(a) Specifically, show that f_{xy} is discontinuous at $(0, 0)$. HINT: Evaluate $f_{xy}(x, y)$ for $(x, y) \neq (0, 0)$. Setting $y = \alpha x$ in that result, show that

$$f_{xy}(x, y) = \frac{1 + 9\alpha^2 - 9\alpha^4 - \alpha^6}{(1 + \alpha^2)^3} \tag{6.1}$$

on each ray $y = \alpha x$ so that the limit of $f_{xy}(x, y)$ as we approach the origin depends on the direction of approach. Thus, conclude that $\lim_{(x,y)\to(0,0)} f_{xy}(x, y) = f_{xy}(0, 0)$ cannot hold.
(b) Proceeding as in part (a), show that $f_{yx}(x, y)$ is discontinuous at $(0, 0)$.
(c) Show that $f_x(x, y)$ given by (4) *is* continuous at $(0, 0)$.
(d) Show that$f_y(x, y)$ given by (6) *is* continuous at $(0, 0)$.

7. If $f_x(x, y)$ and $f_y(x, y)$ exist at (x_0, y_0), does that result imply that f is continuous at (x_0, y_0)? Explain. HINT: Consider the function

$$f(x, y) = \begin{cases} 1, & x = 0 \text{ or } y = 0 \\ 0, & x \neq 0 \text{ and } y \neq 0. \end{cases}$$

*See, for instance, T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957).

13.4 Composite Functions and Chain Differentiation

Let $f(x)$ be differentiable over some interval $X = \{x|\, a < x < b\}$. If $x = x(t)$ is, in turn, a differentiable function of the variable t over some interval $T = \{t|\, \alpha < t < \beta\}$, such that the value $x(t)$ is in X whenever t is in T, then we speak of $f(x) = f(x(t)) \equiv F(t)$ as a **composite function** of t. As noted in a first course in calculus, we can compute dF/dt by the **chain rule**,

$$\frac{dF}{dt} = \frac{df}{dx}\frac{dx}{dt}. \tag{1}$$

EXAMPLE 1. If $f(x) = \sin x$ and $x = t^2$ (over $-\infty < t < \infty$, say), then

$$f(x(t)) = \sin(t^2) \equiv F(t) \qquad \text{and} \qquad \frac{dF}{dt} = \frac{df}{dx}\frac{dx}{dt} = (\cos x)(2t) = 2t\cos(t^2). \quad \blacksquare$$

Just as the chain rule is indispensible in the calculus of functions of a single variable, its extension to functions of several variables is indispensible here. For the case $f(x(t), y(t))$, we have the following basic theorem.

THEOREM 13.4.1 *Chain Rule*
Let $f(x, y)$, $f_x(x, y)$ and $f_y(x, y)$ be continuous at each point of an open region $\mathcal{R}$ in the x, y plane. And let $x = x(t)$, $y = y(t)$ be differentiable functions of t over some open interval T on the t axis, such that the point $(x(t), y(t))$ is in $\mathcal{R}$ whenever t is in T. Then the composite function $f(x(t), y(t)) \equiv F(t)$ is a differentiable function of t for all t in T, and

$$\boxed{\frac{dF}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.} \tag{2}$$

Proof: Let t be in T, and let Δt be small enough so that $t + \Delta t$ is in T as well. Define

$$\Delta x \equiv x(t + \Delta t) - x(t) \qquad \text{and} \qquad \Delta y \equiv y(t + \Delta t) - y(t). \tag{3}$$

Then

$$\begin{aligned}\Delta F &= F(t + \Delta t) - F(t)\\ &= f(x(t + \Delta t), y(t + \Delta t)) - f(x(t), y(t))\\ &= f(x + \Delta x, y + \Delta y) - f(x, y)\\ &= [f(x, y + \Delta y) - f(x, y)] + [f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)]\\ &= f|_P^Q + f|_Q^R,\end{aligned} \tag{4}$$

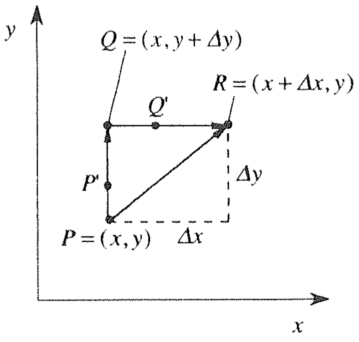


Figure 1. Following constant-coordinate lines.

where P, Q, R are the points shown in Fig. 1.[*] That is, instead of moving from P to R, we break the trip into two parts, from P to Q and then from Q to R, in order to follow constant-coordinate lines.

Next, we apply the mean value theorem of the differential calculus,[†] to each of the terms $f|_P^Q$ and $f|_Q^R$ in (4):

$$\Delta F = \left.\frac{\partial f}{\partial y}\right|_{P'} \Delta y + \left.\frac{\partial f}{\partial x}\right|_{Q'} \Delta x, \tag{5}$$

where Q' is some point between Q and R, and P' is some point between P and Q, as shown in Fig. 1. (This step is justified since it was assumed that f is continuous and that f_x and f_y exist within $\mathcal{R}$.)

Finally, divide (5) by Δt and let $\Delta t \to 0$. As $\Delta t \to 0$, we see from (3) that $\Delta x \to 0$ and $\Delta y \to 0$; hence $Q' \to P$ and $P' \to P$. Thus

$$\lim_{\Delta t \to 0} \frac{\Delta F}{\Delta t} = \lim_{\Delta t \to 0} \left(\left.\frac{\partial f}{\partial x}\right|_{Q'} \frac{\Delta x}{\Delta t} + \left.\frac{\partial f}{\partial y}\right|_{P'} \frac{\Delta y}{\Delta t} \right)$$

becomes[‡]

$$\frac{dF}{dt} = \left(\lim_{Q' \to P} \frac{\partial f}{\partial x} \right) \frac{dx}{dt} + \left(\lim_{P' \to P} \frac{\partial f}{\partial y} \right) \frac{dy}{dt}. \tag{6}$$

And if the partials f_x and f_y are continuous, as assumed, then the limit of f_x as $Q' \to P$ is equal to f_x *at* P, and similarly for f_y. Thus, (6) yields

$$\frac{dF}{dt} = f_x(x,y)\frac{dx}{dt} + f_y(x,y)\frac{dy}{dt}$$

as was to be shown. ■

Notice that the assumed continuity of the partials f_x, f_y is finally called upon in the last step of the proof. Our need for continuity of the partials, at the point in question, is quite reasonable since (2) is essentially an *interpolation* formula, wherein the change in F is computed as a linear combination of the rates of change in the perpendicular x and y directions, for interpolation is viable only if the quantities involved are continuous.

EXAMPLE 2. Let $r(x,y) = x^2 y - e^{2y}$, where

$$x = 3t^2, \qquad y = \sin t. \qquad (1 < t < 4)$$

[*] We use the symbol $(\ )|_P$ to denote $(\ )$ evaluated at P, and $(\ )|_P^Q$ to denote $(\ )|_Q - (\ )|_P$. Incidentally, observe that in the fourth line of (4) we have added and subtracted the quantity $f(x, y + \Delta y)$. This simple idea, the adding and subtracting of a quantity, is often useful.

[†] The **mean value theorem**: If $f(x)$ is continuous over $a \le x \le b$, and $f'(x)$ exists over $a < x < b$, then there is at least one point x_1 between a and b such that $f(b) - f(a) = f'(x_1)(b-a)$.

[‡] Recall from the calculus that $\lim(A+B) = \lim A + \lim B$ and $\lim(AB) = (\lim A)(\lim B)$ if $\lim A$ and $\lim B$ both exist. Thus $\lim(AB+CD) = \lim(AB) + \lim(CD) = (\lim A)(\lim B) + (\lim C)(\lim D)$ if the four limits on the right-hand side exist.

Then $r = r(x(t), y(t)) \equiv R(t)$, and

$$\begin{aligned}\frac{dR}{dt} &= \frac{\partial r}{\partial x}\frac{dx}{dt} + \frac{\partial r}{\partial y}\frac{dy}{dt} \\ &= (2xy)(6t) + (x^2 - 2e^{2y})(\cos t) \\ &= 36t^3 \sin t + (9t^4 - 2e^{2\sin t}) \cos t\end{aligned}$$

for $1 < t < 4$. ▮

EXAMPLE 3. Let $f(u, v) = uv^2$, where

$$u = 3x - y, \qquad v = x^2 y.$$

Then $f(u, v) = f(u(x, y), v(x, y)) \equiv F(x, y)$. To compute $\partial F/\partial x$, say, use the chain rule (2):

$$\begin{aligned}\frac{\partial F}{\partial x} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} \\ &= (v^2)(3) + (2uv)(2xy) \\ &= 3x^4y^2 + 4x^3y^2(3x - y) \\ &= 15x^4y^2 - 4x^3y^3. \end{aligned} \tag{7}$$

COMMENT 1. The chain rule (2) applies even though u and v are functions of more than one variable (x and y) because y is held fixed when we compute $\partial F/\partial x$ so that $u(x, y)$ and $v(x, y)$ are, for the moment, considered as functions of x alone.

COMMENT 2. Is it clear which variables are held fixed when doing the various partial derivatives? There are two sets of variables, $\{x, y\}$ and $\{u, v\}$; $\partial/\partial x$ means the derivative with respect to x with all other variables in the x set (namely, y) fixed, $\partial/\partial u$ means the derivative with respect to u with all other variables in the u set (namely, v) fixed, and so on. ▮

Extension to more than two variables should be evident. Namely, if $f[x_1(t), \ldots, x_n(t)] \equiv F(t)$, then the chain rule (2) becomes

$$\boxed{\frac{dF}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt},} \tag{8}$$

subject to conditions analogous to those stated in Theorem 13.4.1.

Before closing this section we need to discuss the notation that we have been using. If f is a function of x and, in turn, x is a function of t, then we distinguish $f(x)$ from $f(x(t))$ by denoting the latter as $F(t)$. That is, it is necessary to introduce a new function name (F, say) because the f and F functions are different (in general). To illustrate, let $f(x) = \sin x$ and $x = t^2$ (as in Example 1), so then $F(t) = \sin(t^2)$. With $x = 3$, for instance, $f(3) = \sin 3$ whereas $F(3) = \sin 9$, and these values are not the same.

However, beware that in the engineering and science literature this notational distinction is usually not observed. For instance, if $f(x) = \sin x$ and $x = t^2$, then the writer will probably follow with $f(t) = \sin(t^2)$.* As explained above, the latter is incorrect and can cause problems. To illustrate, suppose f is a function of x, y, u, v, and that u, v are, in turn, functions of x and y. Then we should distinguish the function $f(x, y, u, v)$ of x, y, u, v from the function $F(x, y) = f(x, y, u(x, y), v(x, y))$ of x and y. The chain rule gives

$$\frac{\partial F}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}, \tag{9}$$

and similarly for $\partial F/\partial y$. Notice that if we had not introduced a separate function name, F, then the left side of (9) would be $\partial f/\partial x$ and we might cancel the two $\partial f/\partial x$ terms. That step would be incorrect since the one on the left is the partial derivative of $f(\underline{x}, y, u(\underline{x}, y), v(\underline{x}, y))$ with respect to all of the underlined x dependence (with y fixed), whereas the one on the right is the partial derivative of $f(\underline{x}, y, u, v)$ with respect to only the first of the four arguments (the underlined x).

Closure. The key point is the chain rule (2) and its extension to n variables, given by (8). We stress that the chain rule is, essentially, an interpolation formula and, as such, it requires continuity of the partial derivatives [$\partial f/\partial x$ and $\partial f/\partial y$ in (2), and $\partial f/\partial x_1, \ldots, \partial f/\partial x_n$ in (8)].

EXERCISES 13.4

1. Let $f(x, y) = \sin(x^4 + 3y)$, where $x = 5t$ and $y = t^2 + 1$, and denote $f(x(t), y(t)) = F(t)$. Evaluate dF/dt using the chain rule,

$$\frac{dF}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}. \tag{1.1}$$

NOTE: Actually, (1.1) is not the end of the "chain differentiation story," for in computing $\partial f/\partial x$ we set $x^4 + 3y = u$, so that, again applying chain differentiation,

$$\frac{\partial f}{\partial x} = \frac{d}{du}(\sin u)\frac{\partial u}{\partial x} = \text{etc.},$$

and similarly for $\partial f/\partial y$.

2. Let $f(x, y) = e^{xy}$, and denote $f(x(t), y(t)) = F(t)$. Evaluate dF/dt in each case, using the chain rule.

(a) $x = t^2 - 1, \quad y = \sin 3t$
(b) $x = \sqrt{t+1}, \quad \cos t$
(c) $x = t^2, \quad y = 1/(t^2 + 1)$
(d) $x = \ln t, \quad y = t \quad (t > 0)$
(e) $x = \sin t, \quad y = \cos t$
(f) $x = 3t - 1, \quad y = 2t + 5$

3. Let $g(u, v) = \sqrt{u^2 - v}$ and denote $g(u(s), v(s)) = G(s)$. Evaluate dG/ds in each case, using the chain rule. Next, evaluate dG/ds directly; i.e., put $u(s)$ and $v(s)$ into $g(u, v)$, to obtain $G(s)$, and then compute dG/ds. Show that your two answers are the same.

(a) $u = \sin s, \quad v = \cos 6s$ (b) $u = 4s^2, \quad v = e^{-3s}$
(c) $u = s + 1, \quad v = s^3$ (d) $u = \sin 5s, \quad v = 4$
(e) $u = 3s, \quad v = s^2 + 2$ (f) $u = \sin s, \quad v = \cos^2 s$

*Similarly, engineering and science writers typically write $\int_0^x f(x)\,dx$ rather than introducing a new letter for the dummy variable of integration and writing $\int_0^x f(\xi)\,d\xi$. Though use of the former is not likely to lead to an incorrect result, it is important to keep in mind that the x's in $f(x)\,dx$ are not the same as the x in the upper limit. The upper limit x is a fixed endpoint, whereas the x in $f(x)\,dx$ is a dummy variable that varies over the interval of integration.

4. If $r = \sqrt{x^2(t) + y^2(t) + z^2(t)} = R(t)$ is the distance of a particle from the origin of a Cartesian x, y, z coordinate system and t is the time, use the chain rule to determine the radial speed dR/dt in each case, at $t = 2$.

(a) $x = t, \quad y = \sin t, \quad z = 4$
(b) $x = 2t, \quad y = t^2, \quad z = t^3 + 1$
(c) $x = \cos 2t, \quad y = 3, \quad z = \sin 2t$
(d) $x = e^t, \quad y = 4t, \quad z = e^{-t}$
(e) $x = \cos t, \quad y = \sin t, \quad z = 6t$
(f) $x = 3, \quad y = 1 + 2t, \quad z = t^2$

5. (a) Let the Cartesian coordinates of a fish be x, y, z. If the fish swims so that $x = 6t, \quad y = t + 2, \quad z = e^{-t}$, where t is the time, and the temperature distribution in the water is given by

$$T(x, y, z) = \left(5 + \frac{1}{x^2 + y^2 + z^2}\right) e^{-z},$$

determine the time rate of change of temperature experienced by the fish when $t = 0$.
(b) Repeat part (a), with

$$T(x, y, z, t) = \frac{60e^z(1 + 0.1 \sin t)}{x^2 + y^2 + 2}.$$

6. (*Euler's theorem*) $f(x_1, \ldots, x_n)$ is said to be **homogeneous of degree** k if

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^k f(x_1, \ldots, x_n). \tag{6.1}$$

(a) In each case, show whether the function is homogeneous and, if it is, indicate its degree: $f = x^2 + 3xy, \quad g = \ln(x^2 + y^2), \quad h = (x^2 - xy)/(2x + y)$, and $p = x^3 e^{x/2y}$. (Assume that $x^2 + y^2 \neq 0$ in g, that $2x + y \neq 0$ in h, and that $y \neq 0$ in p.)
(b) If $f(u_1, \ldots, u_m)$ is a homogeneous function of $u_1, \ldots, u_m$ of degree p, and $u_1(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n)$ are homogeneous functions of $x_1, \ldots, x_n$ of degree q, show whether f is necessarily a homogeneous function of $x_1, \ldots, x_n$. If so, of what degree?
(c) Show that if $f(x, y, z)$ is homogeneous of degree k, and has partial derivatives $\partial f/\partial x, \ \partial f/\partial y, \ \partial f/\partial z$, then these partial derivatives are homogeneous of degree $k - 1$.
(d) (*Euler's theorem on homogeneous functions*) Prove **Euler's theorem**, that if $f(x_1, \ldots, x_n)$ is homogeneous of degree k and has continuous first-order partial derivatives, then

$$x_1 \frac{\partial f}{\partial x_1} + x_2 \frac{\partial f}{\partial x_2} + \cdots + x_n \frac{\partial f}{\partial x_n} = kf. \tag{6.2}$$

HINT: Differentiate equation (6.1) with respect to λ, then set $\lambda = 1$.

7. Verify Euler's theorem [(6.2) in Exercise 6] for these cases.

(a) $f(x, y) = \sqrt{x^4 + 2y^4} \sin(3x/y), \quad y \neq 0$
(b) $f(x, y) = (x^3 - 2x^2 y + 5y^3)/(x^2 + y^2), \quad x^2 + y^2 \neq 0$

8. Show that $u = f(x + ct) + g(x - ct)$ satisfies the partial differential equation $c^2 u_{xx} - y_{tt} = 0$ (the *wave equation*), where c is a constant and f and g are arbitrary twice-differentiable functions.

9. The differential equation $xy'' + y' + xy = 0$, known as *Bessel's equation of order zero*, has the general solution $y = AJ_0(x) + BY_0(x)$, where J_0 is the *Bessel function of the first kind and order zero*, and Y_0 is the *Bessel function of the second kind and order zero*.

(a) Solve $xy'' + y' + k^2 xy = 0$ in terms of Bessel functions. HINT: Set $x = \alpha t$ [and $y(x) \equiv u(t)$], and choose α so that the differential equation on $u(t)$ is the Bessel equation.
(b) Solve $xy'' + y' + k^2 y = 0$ in terms of Bessel functions. HINT: Set $x = \alpha t^\beta$ [and $y(x) = y(\alpha t^\beta) \equiv u(t)$], and choose α, β so that the new differential equation is the Bessel equation.

13.5 Taylor's Formula and Mean Value Theorem

Just as the Taylor series of a function of a single real variable is indispensible in applied mathematics, so is the Taylor series of a function of several real variables. We expect that you have already studied the former in the calculus but not the latter.

In Section 13.5.1 we review Taylor's formula, Taylor series, and the mean value theorem, and in Section 13.5.2 we extend those results to functions of several real variables.

13.5.1. Taylor's formula and Taylor series for $f(x)$. To develop Taylor's formula about an initial point a, we begin with the identity

$$\int_a^x f'(x)\,dx = f(x) - f(a), \tag{1}$$

where in this discussion it will prove more convenient to forego using different letters for the dummy variables of integration. Solving (1) for $f(x)$,

$$f(x) = f(a) + \int_a^x f'(x)\,dx. \tag{2}$$

Just as (2) holds for $f(x)$, it holds if we replace $f(x)$ by $f'(x)$:

$$f'(x) = f'(a) + \int_a^x f''(x)\,dx. \tag{3}$$

[We obtained (3) by changing f to f' in (2), not by differentiating (2).] Putting this expression for $f'(x)$ into the integral in (2) gives

$$\begin{aligned} f(x) &= f(a) + \int_a^x \left[f'(a) + \int_a^x f''(x)\,dx \right] dx \\ &= f(a) + f'(a)(x-a) + \int_a^x \int_a^x f''(x)\,dx\,dx. \end{aligned} \tag{4}$$

Next, replacing f' in (3) by f'' gives

$$f''(x) = f''(a) + \int_a^x f'''(x)\,dx \tag{5}$$

and putting that expression into the integral in (4) gives

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \int_a^x \int_a^x \left[f''(a) + \int_a^x f'''(x)\,dx \right] dx\,dx \\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\ &\quad + \int_a^x \int_a^x \int_a^x f'''(x)\,dx\,dx\,dx. \end{aligned} \tag{6}$$

Repeating this process (assuming that f is sufficiently differentiable*) gives

$$\boxed{\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\ &\quad + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + R_n(x), \end{aligned}} \tag{7a}$$

*For instance, $\sin x$ is infinitely differentiable (i.e., it admits derivatives of all orders) whereas $x^3 H(x)$ (where H is the Heaviside function) is only twice differentiable (on any interval containing $x = 0$).

where the **remainder term** is

$$R_n(x) = \int_a^x \cdots \int_a^x f^{(n)}(x)\,(dx)^n \tag{7b}$$

and $(dx)^n$ denotes $dx \cdots dx$, n times; (7a) is known as **Taylor's formula**, with the remainder term given in **integral form** by (7b).

To obtain an alternative form for R_n, suppose that $f^{(n)}(x)$ has a minimum value m and a maximum value M on $[a, x]$.* Surely, then,

$$\int_a^x \cdots \int_a^x m\,(dx)^n \leq R_n(x) \leq \int_a^x \cdots \int_a^x M\,(dx)^n \tag{8}$$

or, upon carrying out the integrations,

$$\frac{m}{n!}(x-a)^n \leq R_n(x) \leq \frac{M}{n!}(x-a)^n. \tag{9}$$

If we assume that $f^{(n)}(x)$ is continuous on $[a, x]$, then (it can be shown that) it must take on all values, from its minimum m to its maximum M, over the interval. It therefore follows from (9) that we must be able to express

$$R_n(x) = \frac{f^{(n)}(\xi)}{n!}(x-a)^n \tag{10}$$

for some suitable point ξ in $[a, x]$; (10) gives the **Lagrange form** of R_n, which is often more convenient than the integral form (7b).

If we put (10) into (7a) it appears that the resulting right-hand side is an $(n-1)$th-degree polynomial representation of $f(x)$. Something seems amiss because we know that functions such as $f(x) = e^x$, $\sin x$, and the Bessel function $J_0(x)$ cannot be expressed as finite-degree polynomials. The catch is that ξ, in (10), is not merely a constant, it is a function of the endpoint x. For instance, since ξ is somewhere in $[a, x]$, then it follows that if we let x approach a, then ξ must approach a too; thus, ξ depends upon x. That is, (10) really means

$$R_n(x) = \frac{f^{(n)}(\xi(x))}{n!}(x-a)^n, \tag{11}$$

and (in general) we do not know $\xi(x)$, except that it is somewhere in $[a, x]$. Thus, if we put (10) into (7a) then the resulting right-hand side is not really a polynomial.

Nonetheless, Taylor's formula with Lagrange remainder is valuable because we can write, from (7a),

$$f(x) \approx f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}, \tag{12}$$

*Recall that $[\alpha, \beta]$ denotes the **closed interval** $\alpha \leq x \leq \beta$. Actually, x can be to the right *or* left of the point a, so when we write $[a, x]$ we will mean $[a, x]$ if $x > a$, and $[x, a]$ if $x < a$.

where the error in the approximation is $R_n(x)$, and we can use the form (10) to "bound" that error. Thus, the point is that Taylor's formula (7a) is about *approximation*, the approximation of a given function $f(x)$ by a finite-degree polynomial, with $R_n(x)$ enabling us to obtain a bound on the error thereby incurred.

EXAMPLE 1. Use (7a) to approximate the function $f(x) = e^{-x}$ over $0.7 \le x \le 1.3$, by a third-degree polynomial, and obtain a bound on the error. We obtain, from (7a),

$$e^{-x} = e^{-1} - \frac{e^{-1}}{1!}(x-1) + \frac{e^{-1}}{2!}(x-1)^2 - \frac{e^{-1}}{3!}(x-1)^3 + R_4(x) \tag{13}$$

or

$$e^{-x} \approx e^{-1} - \frac{e^{-1}}{1!}(x-1) + \frac{e^{-1}}{2!}(x-1)^2 - \frac{e^{-1}}{3!}(x-1)^3, \tag{14}$$

where the error incurred in (14) is

$$R_4(x) = \frac{e^{-\xi}}{4!}(x-1)^4 \tag{15}$$

for some ξ in the interval [0.7,1.3]. Even without knowing ξ we can bound $R_4(x)$ as

$$|R_4(x)| \le \frac{e^{-0.7}}{4!}(0.3)^4 = 0.000168 \tag{16}$$

because the greatest value of $e^{-\xi}$ on [0.7,1.3] occurs if $\xi = 0.7$, and the greatest value of $(x-1)^4$ occurs if $x = 0.7$ or 1.3 and is $(0.3)^4$.

Thus, if we confine x to the interval $0.7 \le x \le 1.3$, then the error incurred by the approximation (14) is at most ±0.000168. In fact, calculation reveals that the actual absolute magnitude of the error is 0.000132 at $x = 0.7$, 0 at $x = 1$ and 0.000117 at $x = 1.3$. ∎

Suppose we let $n \to \infty$ in (12). The infinite series that results is called the **Taylor series of f**, about the point a, and we denote that series as TS $f|_a$:

$$\boxed{\text{TS } f|_a \equiv \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j,} \tag{17}$$

where $(x-a)^0 \equiv 1$ even if $x = a$, and $0! \equiv 1$ so that the first term of the series is inevitably $f(a)$, as in (12). For the Taylor series of f to exist, about $x = a$, we need f to be infinitely differentiable at $x = a$ so that the coefficients $f^{(j)}(a)/j!$ exist (for $j = 0, 1, 2, \ldots$).

However, for TS $f|_a$ to be *useful* we need two things: we need the series to *converge* on some x interval, and we need the sum function (i.e., the function to which the series converges) to *equal* $f(x)$ on some x interval I. Then we say that TS $f|_a$ **represents** f on I.

Why do we need to explicitly ask that the sum function equal $f(x)$? Is it possible for TS $f|_a$ to converge, on some interval containing $x = a$, and yet not converge to $f(x)$? Let us see.

EXAMPLE 2. Consider the Taylor series of

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{18}$$

about $x = 0$.* (See Fig. 1.) It can be shown (Exercise 7) that $f'(0) = f''(0) = \cdots = 0$ so that

$$\text{TS } f|_0 = 0 + 0x + 0x^2 + \cdots. \tag{19}$$

Surely the latter series converges for all x, but its sum function is identically zero, not the function $f(x)$ defined in (18). Of course the sum function and $f(x)$ do agree at the point of expansion, $x = 0$, as will always be true, but that is not good enough, and we conclude that the Taylor series of f, about $x = 0$, does not represent f on any interval I even though it converges for all x!

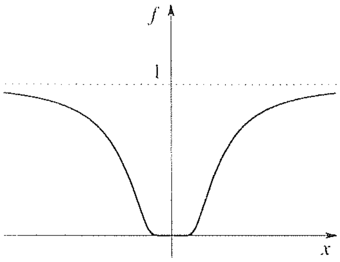


Figure 1. Graph of f.

COMMENT. The failure of TS $f|_0$, as a representation of f, would be reasonable if f were badly behaved at $x = 0$, the point of expansion, but f appears (from Fig. 1) to be well behaved there. To expose the source of the difficulty, we need to examine the behavior of

$$f(z) = \begin{cases} e^{-1/z^2}, & z \neq 0 \\ 0, & z = 0 \end{cases}$$

in the complex $z = x + iy$ plane. As we approach the origin along the real axis (on which $z = x + i0 = x$)

$$\lim_{z\to 0} f(z) = \lim_{x\to 0} e^{-1/x^2} = 0$$

(as can also be seen from Fig. 1), but suppose we approach the origin along the imaginary axis (on which $z = 0 + iy = iy$). In that case

$$\lim_{z\to 0} f(z) = \lim_{y\to 0} e^{-1/(iy)^2} = \lim_{y\to 0} e^{+1/y^2} = \infty,$$

so $f(z)$ is indeed badly behaved at the origin, although that difficulty cannot be observed by looking only along the real axis. The moral of the story is that to fully understand the theory of Taylor series for functions $f(x)$ of a real variable x one must study the theory of Taylor series for functions $f(z)$ of a *complex variable* z. We will do that, but not until Chapter 24. ▮

Meanwhile, we can say the following. If TS $f|_a$ is to represent f, then

$$f(x) = \text{TS } f|_a = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j = \lim_{N\to\infty} \sum_{j=0}^{N} \frac{f^{(j)}(a)}{j!}(x-a)^j \tag{20}$$

*Why do we need to define $f(0) = 0$ separately? Does not $e^{-1/x^2} = e^{-1/0} = e^{-\infty} = 0$ at $x = 0$? No, $e^{-1/0}$ is simply undefined because $1/0$ is undefined. Of course we could define $f(0)$ to be any value we wish but, in making up this example, we chose 0 so that the resulting function would be continuous at $x = 0$.

or, equivalently,

$$\lim_{N\to\infty}\left[f(x)-\sum_{j=0}^{N}\frac{f^{(j)}(a)}{j!}(x-a)^j\right]=0. \tag{21}$$

But, from (7a), we see that (21) is the same as $\lim_{N\to\infty} R_{N+1}(x)=0$ or, equivalently,

$$\lim_{N\to\infty} R_N(x)=0. \tag{22}$$

The condition (22), on some interval I, is both necessary and sufficient for the Taylor series of f to represent f on I. Not only was the condition (22) not satisfied in Example 2, but (7a) became

$$f(x)=0+0+\cdots+0+R_n(x),$$

so that $R_n(x)$ did not tend to zero as $n\to\infty$, it was actually equal to $f(x)$ for each n.

Let us summarize. If we expand a given function $f(x)$ in a Taylor series about some point $x=a$ in the hope that that series will represent f over some x interval, then it does not suffice to test the series for convergence, because it is possible for the series to converge but for $R_n(x)$ not to tend to zero as $n\to\infty$. For if the series converges, but not to $f(x)$, then it does not provide the desired representation of f, as occurred in Example 2. Rather, to show that the Taylor series of a given function f represents f, we need to show that $R_n(x)\to 0$ over some x interval.

EXAMPLE 3. Consider the Taylor series of $f(x)=e^{-x}$ about $x=0$,

$$\text{TS}\, e^{-x}\big|_0 = 1-x+\frac{1}{2!}x^2-\frac{1}{3!}x^3+\cdots. \tag{23}$$

Does that series represent e^{-x} over some x interval so that we can write

$$e^{-x}=1-x+\frac{1}{2!}x^2-\frac{1}{3!}x^3+\cdots? \tag{24}$$

The ratio test (Section 4.2) shows that (23) converges for all x but, as emphasized above, that convergence does not suffice to establish the equality in (24). Everything hinges on the remainder term

$$R_n(x)=\frac{f^{(n)}(\xi)}{n!}(x-a)^n=\frac{(-1)^n e^{-\xi}}{n!}x^n. \tag{25}$$

Since ξ is some point in $[0,x]$ if $x>0$, and in $[x,0]$ if $x<0$, the $e^{-\xi}$ in (25) is at most $e^{|x|}$, so

$$|R_n(x)|\le\frac{e^{|x|}}{n!}|x|^n. \tag{26}$$

Since $|x|^n/n!\to 0$ as $n\to\infty$, for each fixed value of x (no matter how large), it follows that $R_n(x)\to 0$ for every (finite) value of x and, hence, that the equality holds in (24) for all x, for $-\infty<x<\infty$. (See Exercise 8.)

Further, let us explore the convergence of the right-hand side of (24) to e^{-x} in graphical terms. The partial sums of the series are $s_1(x) = 1, \quad s_2(x) = 1 - x, \quad s_3(x) = 1 - x + x^2/2$, and so on, and the first few are shown in Fig. 2. The first partial sum $s_1(x)$ matches the value of e^{-x} at $x = 0$ (i.e., the point of expansion); $s_2(x)$ matches both the value of e^{-x} and its slope at $x = 0$; $s_3(x)$ matches the value of e^{-x}, its slope, and its second derivative at $x = 0$; and so on. It's true that for each fixed n the discrepancy between $s_n(x)$ and e^{-x} becomes infinite as $x \to \infty$, but that fact is irrelevant insofar as convergence is concerned because by the convergence of $s_n(x)$ to e^{-x} we mean that $s_n(x)$ tends to e^{-x}, at a *fixed value of* x, as $n \to \infty$.

Just as a rodeo rider endeavors to stay on the horse, we can imagine the polynomials $s_1(x), \quad s_2(x), \quad \ldots$ endeavoring to "stay on the function:" $s_2(x)$ does a better job than $s_1(x)$, $s_3(x)$ does a better job than $s_2(x)$, and so on, but eventually , as x increases, they all "bite the dust." ▮

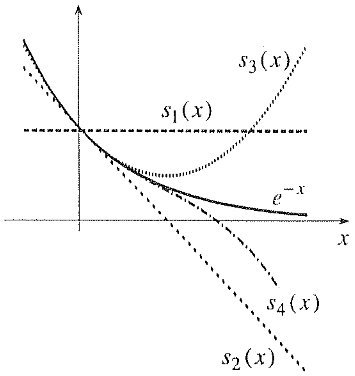


Figure 2. Convergence of the sequence of partial sums to e^{-x}.

The preceding example notwithstanding, it is generally not practical to determine whether or not the Taylor series of a given function f converges to $f(x)$ by determining whether or not $R_n(x) \to 0$ as $n \to \infty$. The reason is that in most examples $R_n(x)$ becomes unwieldy as n increases. Not so in Example 3 because $f^{(n)}(\xi)$ is simply equal to $(-1)^n e^{-\xi}$ if $f(x)$ is e^{-x}, but in general $f^{(n)}(\xi)$, and hence $R_n(x)$, is too unwieldy to bound.

However, functions such as the one in Example 2, whose Taylor series converge, but not to the given function, are rarely encountered in applications. Practically speaking, then, one can merely test the Taylor series of a given function $f(x)$, about a given point $x = a$, for convergence, using Theorem 4.2.2, and assuming (i.e., hoping) that if it converges in $|x - a| < R$ (with the radius of convergence R determined according to the theorem) then it converges to $f(x)$ in that interval and represents $f(x)$ in that interval. This approach is flawed, as discussed above, but should suffice until Taylor series are clarified completely in Chapter 24.

One final application:

EXAMPLE 4. Consider the representation of $f(x) = 1/(1 + x^2)$ by its Taylor series about $x = 0$. To generate the Taylor series of f about $x = 0$ we can use the Taylor series formula (17), but it is easier to recall the geometric series

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots, \tag{27}$$

which holds for $|t| < 1$ (see Exercises 5 and 6 in Section 4.2). That is, identifying t as $-x^2$, we have, from (27),

$$\begin{aligned} f(x) = \frac{1}{1+x^2} = \frac{1}{1-(-x^2)} &= 1 + (-x^2) + (-x^2)^2 + (-x^2)^3 + \cdots \\ &= 1 - x^2 + x^4 - x^6 + \cdots \end{aligned} \tag{28}$$

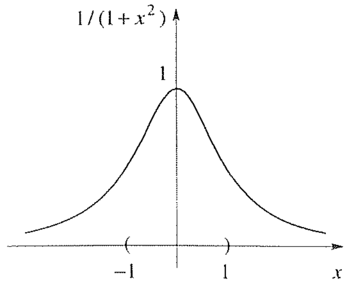


Figure 3. Graph of $f(x) = 1/(1 + x^2)$.

for $|-x^2| < 1$, that is, for $|x| < 1$. From the graph of f (Fig. 3), we wonder why (28) converges only in $|x| < 1$ since $f(x) = 1/(1 + x^2)$ is evidently very well behaved for all

x. As in the case of Example 2, to understand this apparent paradox we need to consider not $f(x) = 1/(1+x^2)$ along the x axis, but $f(z) = 1/(1+z^2)$ in the complex z plane. Specifically, and we don't expect this comment to be completely clear until you study Chapter 24, $f(z) = 1/(1+z^2)$ is singular (becomes infinite) at $z = +i$ and at $z = -i$. Complex Taylor series converge in *disks*, and the Taylor series

$$f(z) = 1 - z^2 + z^4 - z^6 + \cdots \tag{29}$$

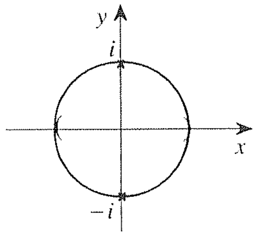


Figure 4. Disk of convergence of (29).

converges in the unit disk (Fig. 4), its radius being limited by the singularities of $f(z)$ at $\pm i$. The intersection of the *disk* $|z| < 1$ with the real axis gives the *interval* of convergence $|x| < 1$ of (28).

Thus, the interval of convergence on the x axis, of the Taylor series of a function $f(x)$, may be limited by singularities of $f(z)$ off of the real axis in the complex z plane. ▮

There is one last point before we extend these ideas to functions of more than one variable. Namely, for the special case where $n = 1$ and where the remainder term $R_1(x)$ is expressed in the Lagrange form (10), Taylor's formula (7a) gives $f(x) = f(a) + f'(\xi)(x-a)$, which result is known as the **mean value theorem.** Let us state that theorem explicitly, for reference.

THEOREM 13.5.1 *Mean Value Theorem for f(x)*
Let f be continuous on the closed interval $[a, x]$* and differentiable on the open interval (a, x). Then there exists a number ξ in (a, x) such that

$$\boxed{f(x) = f(a) + f'(\xi)(x-a).} \tag{30}$$

13.5.2. Extension to functions of more than one variable. You may recall that we have already encountered Taylor series of functions of more than one variable in Example 2 in Section 11.4. We suggest that you review that example either now or when you have completed the present section.

The basic idea behind the Taylor series TS $f|_a$ of a function $f(x)$ is extrapolation: knowing "all about" $f(x)$ at $x = a$ [i.e., knowing $f(a), f'(a), f''(a), \ldots$], can we extrapolate that knowledge and predict what $f(x)$ will be at some other point x? Consider the extrapolation problem in *two* variables. That is, let $f(x, y)$ be defined in an open region $\mathcal{R}$ in the x, y plane, and suppose that all the values of f, f_x, f_y, f_{xx}, f_{xy}, f_{yy}, ... , up to nth order are known at a point (a, b) in $\mathcal{R}$. Can we, from these data, extrapolate from (a, b) and determine the value of f at some other point (x, y) in $\mathcal{R}$?

Often, in mathematics, new ideas are introduced as limits or extensions of old ones. Accordingly, let us try to reduce the stated problem to one involving a single independent variable, since that case is already in hand. Thus, draw a

*As noted earlier, we mean $[a, x]$ if $x > a$, and $[x, a]$ if $x < a$.

straight line from the initial point (a, b) to any desired final point (x_0, y_0), as in Fig. 5, and regard (x, y) as a variable point along that line. The straight line may be parametrized as

$$x = a + (x_0 - a)t, \tag{31a}$$

$$y = b + (y_0 - b)t, \tag{31b}$$

where t is the parameter; for $t = 0$ the point (x, y) coincides with (a, b), and for $t = 1$ it coincides with (x_0, y_0). Then, and this is the point,

$$f(x, y) = f\left(a + (x_0 - a)t, b + (y_0 - b)t\right) \equiv F(t) \tag{32}$$

is a function of the *single variable* t, for $0 \leq t \leq 1$, so if F is sufficiently differentiable we can write Taylor's formula,

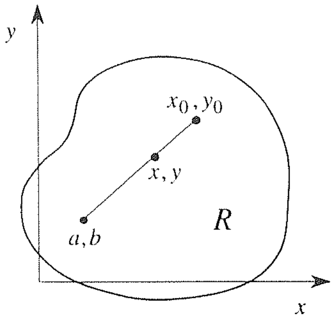


Figure 5. Points of interest.

$$F(t) = F(0) + F'(0)t + \frac{F''(0)}{2!}t^2 + \cdots + \frac{F^{(n-1)}(0)}{(n-1)!}t^{n-1} + R_n(t), \tag{33a}$$

with Lagrange remainder

$$R_n(t) = \frac{F^{(n)}(\tau)}{n!}t^{n+1}, \tag{33b}$$

where τ is some point in $0 \leq \tau \leq 1$. Here we have expanded about $t = 0$ because $t = 0$ corresponds to $x = a$, $y = b$, which is the point in the x, y plane about which we are expanding.

We need to evaluate the coefficients $F(0)$, $F'(0)$, ..., in (33). By chain differentiation,

$$\frac{d}{dt} = \frac{\partial}{\partial x}\frac{dx}{dt} + \frac{\partial}{\partial y}\frac{dy}{dt} = (x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y} \equiv D \tag{34}$$

so

$$F'(0) = \frac{d}{dt}F(t)|_{t=0} = Df(x, y)|_{a,b}, \tag{35a}$$

$$F''(0) = \frac{d}{dt}\frac{d}{dt}F(t)|_{t=0} = D^2 f(x, y)|_{a,b}, \tag{35b}$$

and so on. Putting these expressions into (33) and setting $t = 1$, since $F(1) = f(x_0, y_0)$, gives

$$f(x_0, y_0) = f(a, b) + \frac{1}{1!}Df|_{a,b} + \cdots + \frac{1}{(n-1)!}D^{n-1}f|_{a,b} + R_n(t) \tag{36a}$$

with

$$R_n(t) = \frac{1}{n!}D^n f|_{\xi,\eta}, \tag{36b}$$

where ξ, η are given by the right-hand side of (31) when $t = \tau$ (Fig. 6). Equation (36) is **Taylor's formula** in two variables, with Lagrange remainder. It holds if $f(x, y)$ has continuous partial derivatives (since we used chain differentiation to derive it), through nth order in $\mathcal{R}$, and if the straight line from (a, b) to (x_0, y_0) lies entirely within $\mathcal{R}$.

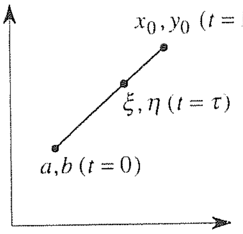


Figure 6. The point (ξ, η).

Remember that the D's in (36) are defined by (34), so

$$\begin{aligned} Df &= [(x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y}]f = (x_0 - a)f_x + (y_0 - b)f_y, \\ D^2 f &= D(Df) = [(x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y}][(x_0 - a)f_x + (y_0 - b)f_y] \\ &= (x_0 - a)^2 f_{xx} + 2(x_0 - a)(y_0 - b)f_{xy} + (y_0 - b)^2 f_{yy}, \end{aligned} \tag{37}$$

and so on. With Df, $D^2 f$, ... in hand, we can write out (36) for any desired n. With $n = 1$, for example, we obtain, with the help of (37),

$$\begin{aligned} f(x_0, y_0) &= f(a, b) + Df|_{\xi,\eta} \\ &= f(a, b) + f_x(\xi, \eta)(x_0 - a) + f_y(\xi, \eta)(y_0 - b). \end{aligned} \tag{38}$$

Notice that x, y were "transition variables" that carried us along the straight line from (a, b) to (x_0, y_0). In the final result, such as (38), x and y no longer appear, only the initial point (a, b) and the final point (x_0, y_0) appear. Since we no longer need to distinguish between x, y and x_0, y_0, we can, for notational simplicity, drop the subscripted zeros and re-express (38) as

$$\boxed{f(x, y) = f(a, b) + f_x(\xi, \eta)(x - a) + f_y(\xi, \eta)(y - b).} \tag{39}$$

The latter is the two-variable version of the **mean value theorem**, which we state, for reference.

THEOREM 13.5.2 *Mean Value Theorem for f(x,y)*
Let $f(x, y)$ and its first-order partial derivatives be continuous in an open region $\mathcal{R}$, and let (a, b) and (x, y) be points in $\mathcal{R}$ such that the straight line joining these points lies entirely within $\mathcal{R}$. Then there exists a point (ξ, η) on that line, between the endpoints, such that

$$f(x, y) = f(a, b) + f_x(\xi, \eta)(x - a) + f_y(\xi, \eta)(y - b). \tag{40}$$

Further, we define the **Taylor series** of $f(x, y)$ about (a, b) as

$$\text{TS } f|_{a,b} = f(a, b) + \frac{1}{1!}Df|_{a,b} + \frac{1}{2!}D^2 f|_{a,b} + \cdots$$

or, with the help of (37) (with x_0 changed to x and y_0 changed to y),

$$\boxed{\begin{aligned}\text{TS}\, f|_{a,b} = f(a,b) &+ \frac{1}{1!}\left[f_x(a,b)(x-a) + f_y(a,b)(y-b)\right] \\ &+\frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b)\right. \\ &\left.+f_{yy}(a,b)(y-b)^2\right] + \cdots.\end{aligned}} \tag{41}$$

We say that $(x-a)^m(y-b)^n$ is of **order** $m+n$, so the dots in (41) denote terms of third order and higher.

EXAMPLE 5. Derive the Taylor series of $f(x,y) = e^{xy}$ about the point $(1,2)$, up to and including second-order terms. Then $f_x = ye^{xy}$, $f_y = xe^{xy}$, $f_{xx} = y^2e^{xy}$, $f_{xy} = (1+xy)e^{xy}$, and $f_{yy} = x^2e^{xy}$, so (41) gives

$$\begin{aligned}\text{TS}\, e^{xy}|_{1,2} &= e^2 + 2e^2(x-1) + e^2(y-2) \\ &\quad + 2e^2(x-1)^2 + 3e^2(x-1)(y-2) + \frac{e^2}{2}(y-2)^2 + \cdots\end{aligned}$$

up to terms of second order. ▮

Extension of these ideas to functions of more than two variables should be straightforward and is left to the exercises.

Before closing this section, let us introduce some notation that will prove useful in what follows. Specifically, if all of the partial derivatives of f, through nth order, are continuous in a region $\mathcal{R}$, then f is said to be **of class C^n** in $\mathcal{R}$.* For example, $f(x,y)$ is of class C^1 in $\mathcal{R}$ if f_x and f_y are continuous in $\mathcal{R}$, and it is of class C^2 in $\mathcal{R}$ if f_x, f_y, f_{xx}, f_{xy}, and f_{yy} are continuous in $\mathcal{R}$. Generally, one omits the words "of class" and simply says that f is C^1 in $\mathcal{R}$, C^2 in $\mathcal{R}$, and so on.

Closure. We begin this section by reviewing and extending the theory of Taylor's formula, Taylor series, and the mean value theorem for functions f of a single variable x. *Taylor's formula* has the advantage that it is the sum of a finite number of terms so convergence of infinite series is not an issue. However, the remainder term, be it in integral form or in Lagrange form, is not known explicitly; for instance, the ξ in (10) is not known. Nevertheless, we can drop the remainder term and *approximate* $f(x)$ by the *polynomial*

$$f(x) \approx f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}$$

*It can be shown that the continuity of the first-order partial derivatives of f imply the continuity of f. Thus, it would be equivalent to say that if f *and* all of the partial derivatives of f, through nth order, are continuous in $\mathcal{R}$, then f is C^n in $\mathcal{R}$.

and use the remainder term $R_n(x)$ to bound the error. *Taylor series*, on the other hand, is an infinite series, so we need to be concerned with whether or not it converges and, if it does, we need to know the x interval over which the convergence is achieved. Further, we emphasized that establishing the convergence of TS $f|_a$ is not the end of the story because it is possible for TS $f|_a$ to converge on an interval I, but not to *represent* f on I. Rather, for TS $f|_a$ to represent f on I it is necessary and sufficient that $R_n(x) \to 0$ on I as $n \to \infty$. Since it is generally difficult to show that $R_n(x) \to 0$ on some interval we are in an awkward position. However, when we study Chapter 24 on complex variable theory we will learn how to establish the interval over which TS $f|_a$ represents $f(x)$, simply by inspection of the function $f(z)$ (where $z = x + iy$). Until then, we suggested merely checking TS $f|_a$ for convergence and not worrying about R_n, on the grounds that in practical applications one rarely encounters functions for which TS $f|_a$ converges without R_n tending to zero as $n \to \infty$.

Computer software. For functions of a single variable, see the computer section at the end of Section 4.2. For multivariable Taylor expansions the *Maple* command is **mtaylor**. For instance, to expand e^{2xy} about $x = 1$, $y = 4$, through terms of second order, enter

```
readlib(mtaylor):
```

Then,

```
mtaylor(exp(2 * x * y), [x = 1, y = 4], 3);
```

and obtain the result

$$e^8 + 8e^8(x-1) + 2e^8(y-4) + 32e^8(x-1)^2 + 18e^8(x-1)(y-4) + 2e^8(y-4)^2$$

EXERCISES 13.5

1. Expand the given function about the indicated point a, through third-order terms. NOTE: $(x-a)^n$ is of nth order.

(a) $e^{-2x}, \quad a = 0$ (b) $e^{-2x}, \quad a = 5$
(c) $e^{-2x}, \quad a = -3$ (d) $\ln x, \quad a = 2$
(e) $1/(1+x^2), \quad a = 1$ (f) $1/(1+x^2), \quad a = -1$
(g) $\sin x, \quad a = 2$ (h) $\cos 2x, \quad a = \pi$
(i) $x(x-1)^2, \quad a = 1$ (j) $x^3(x^4-1)+5, \quad a = 0$

2. Obtain the first four nonvanishing terms in the Taylor series of the given function about $x = 0$.

(a) $1/(1-x^5)$ (b) $1/(2+x^{10})$
(c) $\sin x^{20}$ (d) $\cos x^{20}$

3. Using computer software, generate the Taylor series of the given function, about the indicated point a, through fourth-order terms, and obtain plots of the first five partial sums $[s_1(x)$ through $s_5(x)]$, as well as the given function, over the indicated interval.

(a) $e^x, \quad a = 0, \quad [-2, 2]$
(b) $\cos x, \quad a = 0, \quad [-8, 8]$
(c) $1/(4+x), \quad a = 0, \quad [-3, 3]$
(d) $1/x, \quad a = 5, \quad [1, 9]$
(e) $x^2e^{-x}, \quad a = 2, \quad [-4, 8]$
(f) $(2x-1)^2, \quad a = 0, \quad [-4, 4]$
(g) $1 + x + x^2 + x^3, \quad a = 1, \quad [-3, 5]$
(h) $x^3, \quad a = -1, \quad [-4, 2]$

4. (a) Expanding $\sin x$ about $x = 0$, use (7a) and (10) to show that

$$\sin x = x + R_3(x); \qquad R_3(x) = -\frac{\cos \xi}{3!}x^3,$$

where ξ is some number between 0 and x. Supposing that the interval of interest is $0 \le x \le 0.5$, show that a bound on the error in the approximation $\sin x \approx x$ is

$$|\sin x - x| < 0.021 \quad \text{over} \quad 0 \le x \le 0.5.$$

(b) Similarly, show that

$$\left|\sin x - \left(x - \frac{x^3}{3!}\right)\right| < 0.000261 \quad \text{over} \quad 0 \le x \le 0.5.$$

(c) Similarly, show that

$$\left|\sin x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!}\right)\right| < 0.000011 \quad \text{over} \quad 0 \le x \le 0.5.$$

5. Expanding $\sin x$ about $x = 2$, use (7a) and (10) to show that

$$\sin x = \sin 2 + (\cos 2)(x-2) - \left(\frac{\sin 2}{2}\right)(x-2)^2 + R_3(x),$$

where $R_3(x) = -(\cos\xi)(x-2)^3/3!$ and ξ is some number between 2 and x. Supposing that the interval of interest is $1.5 \le x \le 2.5$, show that

$$\left|\sin x - \left[\sin 2 + (\cos 2)(x-2) - \frac{\sin 2}{2}(x-2)^2\right]\right| < 0.0167$$

over $1.5 \le x \le 2.5$.

6. In the sentence following (9), we speak of assuming the continuity of the derivative $f^{(n)}(x)$. That comment begs a question as to how a derivative can *fail* to be continuous. For instance, suppose $f(x)$ is the "ramp function" $xH(x)$, where H is the Heaviside step function. Then $f'(x) = H(x)$ has a jump discontinuity at $x = 0$. However, at that point the original function $f(x) = xH(x)$ has a "kink" and is not differentiable. Thus, it is hard to imagine how a derivative $f'(x)$ can exist at a point x, yet fail to be continuous there. Nonetheless, the function

$$g(x) = \begin{cases} x^2 \sin\dfrac{1}{x}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{6.1}$$

shows that such behavior is possible because $g'(x)$ exists at $x = 0$ but is not continuous there. The problem that we pose is for you to verify that $g'(x)$ exists at $x = 0$ but is not continuous there. HINT: Evaluate $g'(0)$, and $g'(x)$ for $x \neq 0$, and show that $\lim_{x\to 0} g'(x)$ does not equal $g'(0)$.

7. In Example 2, we stated that $f'(0) = f''(0) = \cdots = 0$. Verify that

(a) $f'(0) = 0$ (b) $f''(0) = 0$ (c) $f'''(0) = 0$

HINT: It will be necessary fo fall back on the limit-of-the-difference-quotient definition of derivative.

8. Below (26), we stated that $|x|^n/n! \to 0$ as $n \to \infty$. Prove that claim. HINT: Use **Stirling's formula**,

$$\boxed{n! \sim \sqrt{2\pi n}\, n^n e^{-n} \quad \text{as} \quad n \to \infty.} \tag{8.1}$$

9. (*Tangent plane*) Let a surface S be defined by an equation $f(x, y, z) = 0$ and let (a, b, c) be a point on S. The tangent plane to S, at (a, b, c), is obtained by replacing $f(x, y, z)$ in $f(x, y, z) = 0$ by its linear approximation $f(a, b, c)$ $+f_x(a, b, c)(x-a)$ $+f_y(a, b, c)(y-b)$ $+f_z(a, b, c)(z-c)$ [i.e., by the first three terms of its Taylor series about (a, b, c)]. Obtain the tangent plane at $(1, 3, -2)$:

(a) $x^3yz = -6$ (b) $xyz = -6$
(c) $3x^2 + y^2 + z^2 = 16$ (d) $\sin(x - y - z) = 0$
(e) $\sin(x^2 + y^2 + z) = \sin 8$ (f) $x^4 + y^2 + z^4 = 26$

10. The essential idea, in our derivation of the two-variable Taylor series (41), was the reduction to a Taylor series in *one* variable by introducing a parametrized line, given by (31), from the initial point (a, b) to the final point (x_0, y_0). Actually, *any* parametrized curve between those two points should give the same final result (41); we chose the straight line (31) as the simplest. In place of (31), use the parametrization

$$x = a + (x_0 - a)t, \qquad y = b + (y_0 - b)t^2, \tag{10.1}$$

which is a parabola, and show that you obtain the same final result (41).

11. (*Expanding in one variable at a time*) An easy way to derive Taylor series in more than one variable is to expand in one variable at a time. Use this method to expand the given function about the given point, through third-order terms.

(a) $x^5y^4 - y$, (1,2) (b) e^{xy}, (1,2)
(c) $\sin(3xy)$, $(1, -1)$ (d) $1/(x^2 + y^2)$, (2,1)
(e) $1/(x^2 - y^2)$, (3,1) (f) $1/(1 + x^2y^2)$, (1,1)

12. (a) – (f) Use computer software to obtain the Taylor series in the corresponding part of exercise 11.

13.6 Implicit Functions and Jacobians

13.6.1. Implicit function theorem. An equation

$$f(x,y)=0 \tag{1}$$

is said to constitute a **relation** on x and y. In (1), x and y have the same status; they are independent variables. Depending on the context, it may be desirable to change our point of view and regard (1) as implicitly defining y as a function of x (or vice versa). In that case we re-express (1) as

$$f(x,y(x))=0, \tag{2}$$

because x is now the independent variable and y is the dependent variable.

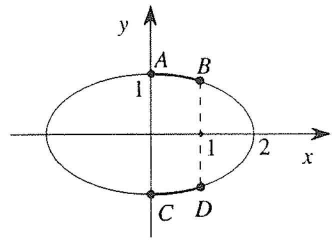


Figure 1. The ellipse $x^2+4y^2-4=0$.

EXAMPLE 1. Consider the relation

$$x^2+4y^2-4=0, \tag{3}$$

satisfied by the points on the ellipse shown in Fig. 1. That (3) is an implicit definition of a function $y(x)$ is clear because we can solve (3) for y, by algebra, as

$$y=\pm\sqrt{1-\left(\frac{x}{2}\right)^2}. \tag{4}$$

If we are interested in the interval $0\le x\le 1$, for instance, then we see that there are actually two continuous functions $y(x)$, the positive square root corresponding to the upper branch AB in Fig. 1, and the negative square root corresponding to the lower branch CD. If, instead, we are interested in the interval $3\le x\le 8$, say, then there is *no* (real valued) function $y(x)$ implied by (3). ▮

In Example 1 we were able to solve (3) for $y(x)$ by simple algebra. In other cases we are not as fortunate. For instance, $2xy+\sin y=3$ is a transcendental equation* and cannot be solved explicitly for y. Thus, there is the following important question. Given an x,y pair (x_0,y_0) satisfying the relation (1), does there exist an **implicit function** $y(x)$ [i.e., a function $y(x)$ implicitly defined by (1)] and, if so, is it unique? Sufficient conditions for that to be true are given by the **implicit function theorem**, which we state without proof.

THEOREM 13.6.1 *Implicit Function Theorem*
Let $f(x,y)=0$ be satisfied by a pair of real numbers x_0,y_0 so that $f(x_0,y_0)=0$, and suppose that $f(x,y)$ is C^1 in some neighborhood of (x_0,y_0) with

$$\boxed{\frac{\partial f}{\partial y}(x_0,y_0)\neq 0.} \tag{5}$$

*In Section 8.2 we define an equation as **algebraic** if it is a finite-degree polynomial equation and **transcendental** otherwise.

Then $f(x, y) = 0$ uniquely implies a function $y(x)$ in some neighborhood N of x_0 such that $y(x_0) = y_0$, where $y(x)$ is differentiable in N.*

Observe that the implicit function theorem is a local one. That is, it assures us that there exists a unique differentiable function $y(x)$ in some neighborhood of x_0, but it does not tell us how large that neighborhood is.

EXAMPLE 2. Consider again the relation (3) in Example 1,

$$f(x, y) = x^2 + 4y^2 - 4 = 0. \tag{6}$$

Noticing that the latter is satisfied by $x = 1$ and $y = -\sqrt{3}/2$, say, let (x_0, y_0) be $(1, -\sqrt{3}/2)$. Clearly, $f_x(x, y) = 2x$ and $f_y(x, y) = 8y$ are continuous in some neighborhood of $(1, -\sqrt{3}/2)$, because they are continuous everywhere in the x, y plane. Thus, f is C^1 in some neighborhood of $(1, -\sqrt{3}/2)$. Further, $f_y(1, -\sqrt{3}/2) = -4\sqrt{3} \neq 0$, so Theorem 13.6.1 tells us that there is, indeed, a unique differentiable function $y(x)$ through the point $(1, -\sqrt{3}/2)$. The graph of that function is the lower half of the ellipse (Fig. 2), over $-2 < x < 2$.

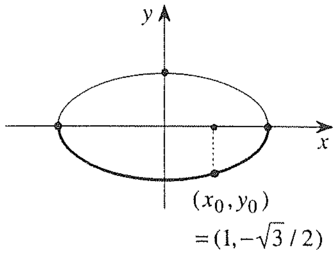


Figure 2. Implicit function $y(x)$ through $(1, -\sqrt{3}/2)$.

COMMENT. If we select (x_0, y_0) to be at one of the two ends, say $(2, 0)$, then the theorem would provide no information since $f_y(2, 0) = (8)(0) = 0$ so that condition (5) is not satisfied. ∎

Whereas Example 2 served to illustrate the implicit function theorem, it did not reveal the power of that theorem because it was such a simple case.

EXAMPLE 3. Consider the relation

$$(y - 2x)e^y - x^2 + 1 = 0. \tag{7}$$

We see by inspection that (7) is satisfied by $x = 1$ and $y = 2$ so let us take $(x_0, y_0) = (1, 2)$. Does there exist a function $y(x)$, implied by (7), through that point? The derivatives

$$f_x = -2e^y - 2x \qquad \text{and} \qquad f_y = (y - 2x + 1)e^y$$

are continuous everywhere in the x, y plane and

$$f_y(1, 2) = e^2 \neq 0$$

so, according to Theorem 13.6.1, there does exist such an implicit function $y(x)$. ∎

How to obtain that function $y(x)$ is another matter. Equation (7) is a transcendental equation in y and cannot be solved for y in closed form. Giving up on the

*$f(x, y)$ being C^1 means that the first-order partial derivatives f_x and f_y are continuous; see Section 13.5, below Example 5. Also, note that the neighborhood N of x_0 is an open *interval* on the x axis, whereas the neighborhood of (x_0, y_0) is an open *disk* in the x, y plane.

hope of obtaining a closed form solution, perhaps we can at least develop the Taylor series of $y(x)$ about x_0 (i.e., an "open" form solution).

Turning from Example 3 to the general case

$$f(x, y(x)) = 0, \tag{8}$$

suppose that we know an x_0 and a y_0 such that $f(x_0, y_0) = 0$, and let us seek the Taylor series of y about x_0,

$$y(x) = y(x_0) + y'(x_0)(x - x_0) + \frac{y''(x_0)}{2!}(x - x_0)^2 + \cdots. \tag{9}$$

The first coefficient, $y(x_0) = y_0$, is already known, but we need to evaluate $y'(x_0)$, $y''(x_0)$, and so on. Applying the chain rule to (8) gives

$$\frac{d}{dx} f(x, y(x)) = f_x + f_y y' = 0 \tag{10}$$

so

$$y'(x) = -\frac{f_x(x, y(x))}{f_y(x, y(x))}. \tag{11}$$

[It is interesting that (10) is a linear algebraic equation for y' even if (8) is a transcendental equation in y.] Continuing, differentiate (11) to find y'':

$$y'' = -\frac{f_{xx} + f_{xy}y'}{f_y} + \frac{f_x(f_{yx} + f_{yy}y')}{f_y^2}. \tag{12}$$

Using (11) for y', and assuming that $f_{xy} = f_{yx}$, (12) can be expressed as

$$y'' = \frac{2f_x f_y f_{xy} - f_x^2 f_{yy} - f_y^2 f_{xx}}{f_y^3}. \tag{13}$$

Similarly for y''', and so on. It remains merely to evaluate the expressions for y', y'', ... at (x_0, y_0).

Now we can appreciate the significance of the condition (5) in Theorem 13.6.1, for it is evident that increasingly higher powers of f_y build up in the denominators of y', y'', ... so that if $f_y(x_0, y_0) = 0$, then these derivatives fail to exist and, consequently, the Taylor series (9) fails to exist.

EXAMPLE 4. Let us return to Example 3 and obtain the Taylor series of $y(x)$ about $x_0 = 1$. Rather than merely evaluating the right-hand sides of (11) and (13), it might be more instructive to carry out steps (10)–(13) for the given equation

$$(y - 2x)e^y - x^2 + 1 = 0. \tag{14}$$

First, differentiating (14) with respect to x gives

$$(y' - 2)e^y + (y - 2x)e^y y' - 2x = 0$$

so

$$y' = \frac{2 + 2xe^{-y}}{y - 2x + 1}. \tag{15}$$

Next, differentiation of (15) gives

$$y'' = \frac{(2 - 2xy')e^{-y}}{y - 2x + 1} + \frac{(2 + 2xe^{-y})(-1)(y' - 2)}{(y - 2x + 1)^2}, \tag{16}$$

and so on. Evaluating (15) and (16) at $(1, 2)$ gives $y'(1, 2) = 2 + 2e^{-2}$ and $y''(1, 2) = -6e^{-2} - 8e^{-4}$ so

$$\begin{aligned} y(x) &= 2 + (2 + 2e^{-2})(x - 1) + \frac{1}{2!}(-6e^{-2} - 8e^{-4})(x - 1)^2 + \cdots \\ &= 2 + 2.271(x - 1) - 0.479(x - 1)^2 + \cdots \end{aligned}$$

is the desired Taylor series. ▮

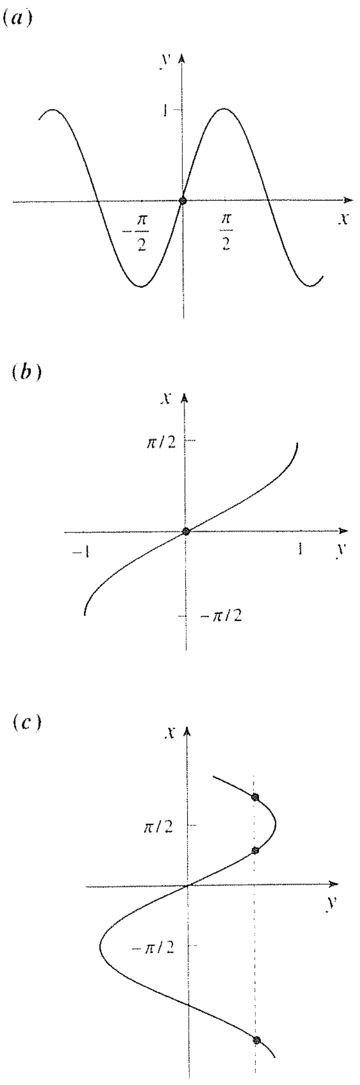


Figure 3. The inverse function $x = \sin^{-1}y$.

Before extending these ideas to the case of two or more independent variables and two or more dependent variables, let us observe that **inverse functions** are but a special case of implicit functions.

EXAMPLE 5. Given the function

$$y = \sin x \tag{17}$$

on $-\infty < x < \infty$, is there an inverse function $x(y)$ through the point $x = 0, y = 0$? Yes, known as $\sin^{-1}y$, its graph is as shown in Fig. 3b and is obtained from Fig. 3a simply by reversing the horizontal and vertical axes. The x interval is limited to $-\pi/2 < x < \pi/2$ because if it were made broader then multivaluedness would result (Fig. 3c), whereas functions are required to be single valued.

Seeking an inverse function $x(y)$ is equivalent to re-expressing (17) as

$$f(x, y) = y - \sin x = 0 \tag{18}$$

and seeking an implicit function $x(y)$. Since the roles of x and y are reversed here, compared to their roles in this section prior to this example, let us re-express (18) as $f(y, x) = y - \sin x = 0$. Since $f_x = -\cos x$ and $f_y = 1$ are continuous everywhere in the x, y plane and $f_x = -\cos x = -1 \neq 0$ at $(x_0, y_0) = (0, 0)$, Theorem 13.6.1 assures us that there does exist a unique differentiable implicit function $x(y)$ (namely, the inverse function $\sin^{-1}y$) in some neighborhood of y_0 (namely, $-1 < y < 1$). ▮

13.6.2. Extension to multivariable case.

Next, consider *two* relations,

$$\begin{aligned} f(x, y, u, v) &= 0, \\ g(x, y, u, v) &= 0, \end{aligned} \tag{19}$$

on x, y, u, v. Supposing that we know a point (x_0, y_0, u_0, v_0) in four-dimensional x, y, u, v space such that $f(x_0, y_0, u_0, v_0) = 0$ and $g(x_0, y_0, u_0, v_0) = 0$, we wonder if (19) implicitly defines functions $u(x, y), v(x, y)$ throughout some neighborhood of x_0, y_0. To explore that possibility, let us re-express (19) as

$$\begin{aligned} f(x, y, u(x, y), v(x, y)) &= 0, \\ g(x, y, u(x, y), v(x, y)) &= 0 \end{aligned} \tag{20}$$

and seek Taylor series of u and v,

$$\begin{aligned} u(x, y) &= u(x_0, y_0) + u_x(x_0, y_0)(x - x_0) + u_y(x_0, y_0)(y - y_0) + \cdots, \\ v(x, y) &= v(x_0, y_0) + v_x(x_0, y_0)(x - x_0) + v_y(x_0, y_0)(y - y_0) + \cdots, \end{aligned} \tag{21}$$

about x_0, y_0. We already know $u(x_0, y_0) = u_0$ and $v(x_0, y_0) = v_0$, but we need to compute $u_x(x_0, y_0)$, $u_y(x_0, y_0), \dots$, and $v_x(x_0, y_0)$, $v_y(x_0, y_0), \dots$. To find u_x and v_x, apply the chain rule to (20):

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y, u(x, y), v(x, y)) &= f_x + f_u u_x + f_v v_x = 0, \\ \frac{\partial}{\partial x} g(x, y, u(x, y), v(x, y)) &= g_x + g_u u_x + g_v v_x = 0 \end{aligned} \tag{22}$$

or

$$\begin{aligned} f_u u_x + f_v v_x &= -f_x, \\ g_u u_x + g_v v_x &= -g_x. \end{aligned} \tag{23}$$

Solving (23) by Cramer's rule (Section 10.6.3 or Appendix B) gives

$$u_x = \frac{\begin{vmatrix} -f_x & f_v \\ -g_x & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_x & f_v \\ g_x & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}, \tag{24a}$$

$$v_x = \frac{\begin{vmatrix} f_u & -f_x \\ g_u & -g_x \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_u & f_x \\ g_u & g_x \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}. \tag{24b}$$

Similarly, to find u_y and v_y take $\partial/\partial y$ of (20) and obtain

$$u_y = \frac{\begin{vmatrix} -f_y & f_v \\ -g_y & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_y & f_v \\ g_y & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}, \tag{25a}$$

$$v_y = \frac{\begin{vmatrix} f_u & -f_y \\ g_u & -g_y \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_u & f_y \\ g_u & g_y \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}. \tag{25b}$$

These results are analogous to those obtained above for the Taylor series expansion of a function $y(x)$ that is implicitly defined by $f(x, y(x)) = 0$. There, we saw from (11) – (13) that the coefficients $y'(x_0)$, $y''(x_0)$, ... in (9) exist if $f_y(x_0, y_0) \neq 0$, and fail to exist if $f_y(x_0, y_0) = 0$. From (24) and (25) we see, analogously, that the derivatives u_x, v_x, u_y, v_y needed in the Taylor series expansions (21) exist if the *determinant*

$$\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix} \tag{26}$$

does not vanish (i.e., equal zero) at x_0, y_0, u_0, v_0, and that they fail to exist if the determinant does vanish at that point. Similarly, we would find that the same determinant condition applies for the existence of the higher-order partial derivatives needed in (21).

Thus, just as the condition

$$\left.\frac{\partial f}{\partial y}\right|_{x_0, y_0} \neq 0 \tag{27}$$

is crucial [recall (5)] if

$$f(x, y) = 0 \tag{28}$$

is to imply the existence of an implicit function $y(x)$, the condition

$$\left.\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}\right|_{x_0, y_0, u_0, v_0} \neq 0 \tag{29}$$

is crucial if

$$f(x, y, u, v) = 0, \tag{30a}$$
$$g(x, y, u, v) = 0 \tag{30b}$$

are to imply the existence of implicit functions $u(x, y)$ and $v(x, y)$. If we realize that the $\partial f/\partial y$ in (27) is actually a "one-by-one determinant," then we can see how the condition (27) generalizes to the nonvanishing of a two-by-two determinant when there are two dependent variables and two independent variables and, similarly, to the nonvanishing of an n-by-n determinant when there are n dependent variables and n independent variables. Thus, the generalized version of Theorem 13.6.1 is as follows.

THEOREM 13.6.2 *Implicit Function Theorem, Multivariable Case*
Let the system of n equations

$$\begin{aligned} f_1(x_1, \dots, x_n, u_1, \dots, u_n) &= 0 \\ &\vdots \\ f_n(x_1, \dots, x_n, u_1, \dots, u_n) &= 0 \end{aligned} \tag{31}$$

be satisfied by the real numbers $x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0}$, and suppose that the functions $f_1(x_1, \ldots, x_n, u_1, \ldots, u_n)$ through $f_n(x_1, \ldots, x_n, u_1, \ldots, u_n)$ are C^1 in some neighborhood of $(x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0})$, with

$$\boxed{\left.\left|\begin{matrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} & \cdots & \dfrac{\partial f_1}{\partial u_n} \\ \vdots & & & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \dfrac{\partial f_n}{\partial u_2} & \cdots & \dfrac{\partial f_n}{\partial u_n} \end{matrix}\right|\right|_{x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0}} \neq 0.} \tag{32}$$

Then (31) uniquely implies functions $u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)$ in some neighborhood N of $(x_{1_0}, \ldots, x_{n_0})$ such that $u_1(x_{1_0}, \ldots, x_{n_0}) = u_{1_0}, \ldots, u_n(x_{1_0}, \ldots, x_{n_0}) = u_{n_0}$, where $u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)$ are C^1 in N.

Although we motivated Theorem 13.6.2, above, we did not prove it.*

EXAMPLE 6. Consider the familiar change of variables from the Cartesian x, y coordinates to the polar coordinates r, θ (Fig. 4),

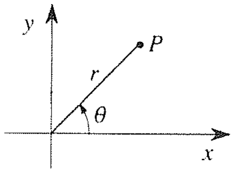


Figure 4. Polar coordinates.

$$x = r\cos\theta, \tag{33a}$$

$$y = r\sin\theta \tag{33b}$$

or, equivalently,

$$\begin{aligned} f_1(x, y, r, \theta) &= x - r\cos\theta = 0, \\ f_2(x, y, r, \theta) &= y - r\sin\theta = 0. \end{aligned} \tag{34}$$

Do these relations define implicit functions $r(x, y)$ and $\theta(x, y)$ throughout some neighborhood of any given point P (Fig. 4)? Put differently, if (33) gives x and y as functions of r and θ, do those relations implicitly define inverse functions $r(x, y)$ and $\theta(x, y)$?

To answer that queston using Theorem 13.6.2 let us identify x, y, r, θ as x_1, x_2, u_1, u_2, respectively. (Whether we identify x as x_1 and y as x_2, or vice versa, will not matter; similarly for r, θ and u_1, u_2.) All first-order partial derivatives of f_1 and f_2 are continuous everywhere so f_1 and f_2 are C^1 everywhere. Further,

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial r} & \dfrac{\partial f_1}{\partial \theta} \\ \dfrac{\partial f_2}{\partial r} & \dfrac{\partial f_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix} = -r \tag{35}$$

is nonzero everywhere except at the origin ($r = 0$). From Theorem 13.6.2 it follows that inverse functions $r(x, y)$ and $\theta(x, y)$ do exist (and are unique and of class C^1) in some neighborhood of any given point P (Fig. 4) except if P is at the origin, in which case the

*See, for instance, I. S. Sokolnikoff, *Advanced Calculus* (New York: McGraw-Hill, 1939), Chapter 12, or W. Maak, *Modern Calculus* (New York: Holt, Rinehart and Winston, 1963), Chapter 9.

theorem gives no information. This singular result is not surprising since there is a problem at the origin; as we can see from the figure, θ is *not defined* there. ∎

13.6.3. Jacobians. The determinant in (32) is known as the **Jacobian** of $f_1, \ldots, f_n$ with respect to $u_1, \ldots, u_n$ and is denoted either by the notation $\dfrac{\partial(f_1, \ldots, f_n)}{\partial(u_1, \ldots, u_n)}$ or as $J(u_1, \ldots, u_n)$:

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} & \cdots & \dfrac{\partial f_1}{\partial u_n} \\ \vdots & & & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \dfrac{\partial f_n}{\partial u_2} & \cdots & \dfrac{\partial f_n}{\partial u_n} \end{vmatrix} \equiv \frac{\partial(f_1, \ldots, f_n)}{\partial(u_1, \ldots, u_n)} \quad \text{or} \quad J(u_1, \ldots, u_n) \tag{36}$$

and is a function of $u_1, \ldots, u_n$.

Jacobian determinants arise not only in the Implicit Function Theorem but also in the evaluation of the various partial derivatives of implicit functions. For instance, consider the functions $u(x, y)$ and $v(x, y)$ defined implicitly by (20). The steps (22)–(25), which we urge you to review, give these results for the partial derivatives of u and v with respect to x and y:

$$u_x = -\frac{\dfrac{\partial(f,g)}{\partial(x,v)}}{\dfrac{\partial(f,g)}{\partial(u,v)}}, \qquad u_y = -\frac{\dfrac{\partial(f,g)}{\partial(y,v)}}{\dfrac{\partial(f,g)}{\partial(u,v)}}, \tag{37a,b}$$

$$v_x = -\frac{\dfrac{\partial(f,g)}{\partial(u,x)}}{\dfrac{\partial(f,g)}{\partial(u,v)}}, \qquad v_y = -\frac{\dfrac{\partial(f,g)}{\partial(u,y)}}{\dfrac{\partial(f,g)}{\partial(u,v)}}. \tag{38a,b}$$

EXAMPLE 7. Consider, once again, the change of variables from x, y to r, θ, given by (33). We will refer to x_r, x_θ, y_r and y_θ as the "forward" partial derivatives, since x and y are given explicitly by (33) as functions of r and θ. The forward derivatives follow readily from (33) as

$$x_r = \cos\theta, \qquad x_\theta = -r\sin\theta, \qquad y_r = \sin\theta, \qquad y_\theta = r\cos\theta, \tag{39}$$

but to evaluate the reverse derivatives is harder; we can carry out the steps (22)–(25) or we can simply use the final results, reproduced (in terms of Jacobians) in (37) and (38). To use (37) and (38), let u be r and let v be θ. Then

$$\begin{aligned} f(x, y, r, \theta) &= x - r\cos\theta = 0, \\ g(x, y, r, \theta) &= y - r\sin\theta = 0 \end{aligned} \tag{40}$$

so (37) gives

$$r_x = -\frac{\begin{vmatrix} f_x & f_\theta \\ g_x & g_\theta \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} 1 & r\sin\theta \\ 0 & -r\cos\theta \end{vmatrix}}{\begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix}} = \cos\theta, \tag{41a}$$

$$r_y = -\frac{\begin{vmatrix} f_y & f_\theta \\ g_y & g_\theta \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} 0 & r\sin\theta \\ 1 & -r\cos\theta \end{vmatrix}}{\begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix}} = \sin\theta, \tag{41b}$$

and (38) gives

$$\theta_x = -\frac{\begin{vmatrix} f_r & f_x \\ g_r & g_x \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} -\cos\theta & 1 \\ -\sin\theta & 0 \end{vmatrix}}{r} = -\frac{\sin\theta}{r}, \tag{42a}$$

$$\theta_y = -\frac{\begin{vmatrix} f_r & f_y \\ g_r & g_y \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} -\cos\theta & 0 \\ -\sin\theta & 1 \end{vmatrix}}{r} = \frac{\cos\theta}{r}. \tag{42b}$$

Of course these results can be expressed in terms of x, y if we wish. For instance, $\theta_y = r\cos\theta/r^2 = x/(x^2+y^2)$.

COMMENT. In this case we could have solved (33) explicitly for $r(x, y)$ and $\theta(x, y)$, for squaring and adding those equations gives r and dividing one by the other gives θ:

$$r = \sqrt{x^2+y^2}, \qquad \theta = \tan^{-1}\frac{y}{x}, \tag{43}$$

from which we can compute the "forward" derivatives r_x, r_y, θ_x, and θ_y. However, we cannot always solve for the inverse functions analytically. Hence, we have used this example to illustrate the use of the Jacobians in solving for the reverse derivatives. ▮

Why all the fuss? Could we not have solved for the reverse derivatives by finding the forward derivatives (which is simple) and "turning them upside down"? For instance, knowing that $\partial x/\partial r = \cos\theta$, could we not say that $\partial r/\partial x = \frac{1}{\partial x/\partial r} = \frac{1}{\cos\theta}$? Evidently not, since (41a) gave $\partial r/\partial x = \cos\theta$, not $1/\cos\theta$.*

To understand this important point, observe that if $x = x(u)$ has an inverse function $u = u(x)$, then du/dx *does* equal $(dx/du)^{-1}$ or, equivalently,

$$\frac{du}{dx}\frac{dx}{du} = 1. \tag{44}$$

*The fact that, besides not being numerical inverses of each other, r_x and x_r are actually equal to each other is just a coincidence.

For instance, if $x = u^2$ then $dx/du = 2u$ and $du/dx = d(\sqrt{x})/dx = 1/(2\sqrt{x}) = 1/(2u) = (dx/du)^{-1}$. However, if

$$x = x(u,v) \quad \text{and} \quad y = y(u,v) \tag{45a}$$

have inverse functions

$$u = u(x,y) \quad \text{and} \quad v = v(x,y), \tag{45b}$$

then (in general)

$$\frac{\partial u}{\partial x}\frac{\partial x}{\partial u} \neq 1, \quad \frac{\partial u}{\partial y}\frac{\partial y}{\partial u} \neq 1, \quad \frac{\partial v}{\partial x}\frac{\partial x}{\partial v} \neq 1, \quad \frac{\partial v}{\partial y}\frac{\partial y}{\partial v} \neq 1. \tag{46}$$

Rather, what follows from (45) is the *Jacobian* generalization of (44),

$$\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(u,v)} = 1, \tag{47}$$

rather than (46). Similarly, if $x_1, \ldots, x_n$ are functions of $u_1, \ldots, u_n$, and vice versa, then

$$\boxed{\frac{\partial(u_1,\ldots,u_n)}{\partial(x_1,\ldots,x_n)}\frac{\partial(x_1,\ldots,x_n)}{\partial(u_1,\ldots,u_n)} = 1.} \tag{48}$$

Summarizing, while it is *not* true (in general) that the individual partial derivatives are numerical inverses of each other [e.g., $\dfrac{\partial u_2}{\partial x_3} \neq \left(\dfrac{\partial x_3}{\partial u_2}\right)^{-1}$], it *is* true that the *Jacobians* are numerical inverses of each other:

$$\frac{\partial(u_1,\ldots,u_n)}{\partial(x_1,\ldots,x_n)} = \left(\frac{\partial(x_1,\ldots,x_n)}{\partial(u_1,\ldots,u_n)}\right)^{-1}. \tag{49}$$

Rather than prove the general case (48), let it suffice to prove (47), where $n = 2$. We begin by writing out

$$\begin{aligned}
\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(u,v)} &= \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} \\
&= (u_x v_y - u_y v_x)(x_u y_v - x_v y_u) \\
&= u_x v_y x_u y_v - u_y \underline{v_x x_u} y_v \\
&\quad - \underline{u_x} v_y \underline{x_v} y_u + u_y v_x x_v y_u.
\end{aligned} \tag{50}$$

Next, use the chain derivative results*

$$\frac{\partial v}{\partial u} = 0 = \frac{\partial v}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial v}{\partial y}\frac{\partial y}{\partial u} \quad \text{so} \quad v_x x_u = -v_y y_u,$$

$$\frac{\partial u}{\partial v} = 0 = \frac{\partial u}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial v} \quad \text{so} \quad u_x x_v = -u_y y_v$$

*In case it is not clear why $\partial v/\partial u = 0$, for instance, think of $\{x, y\}$ and $\{u, v\}$ as two "families" of variables. By $\partial/\partial u$ we mean differentiation with respect to u, holding all other members of the u family (namely, v) fixed. Because v is regarded as fixed, $\partial v/\partial u = 0$.

to replace the underlined products in (50). Then the right-hand side of (50) becomes

$$\begin{aligned}\text{RHS} &= u_x v_y x_u y_v + u_y y_v v_y y_u \\ &\qquad + v_y y_u u_y y_v + u_y v_x x_v y_u \\ &= y_v v_y (u_x x_u + u_y y_u) + y_u u_y (v_y y_v + v_x x_v) \\ &= y_v v_y u_u + y_u u_y v_v \\ &= y_v v_y + y_u u_y = \frac{\partial y}{\partial y} \\ &= 1, \end{aligned} \tag{51}$$

as claimed. Naturally, we assumed that all the first-order partial derivatives are continuous so that the chain rule could be used.

13.6.4. Application to change of variables. It sometimes happens that it is convenient to change from one set of independent variables to another, and that step involves the concepts discussed above.

To illustrate, suppose that we wish to determine the temperature distribution $T(x, y)$ in a semicircular plate, with the temperature maintained at 100° along the circular edge and at 0° along the straight edge, as depicted in Fig. 5. It turns out, as we will find in Chapter 16, that $T(x, y)$ must satisfy the *Laplace equation*, namely, the partial differential equation

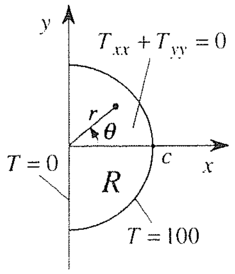


Figure 5. Heat conduction problem.

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \tag{52}$$

in the semicircular region $\mathcal{R}$, together with the boundary conditions that $T = 100$ on the circular edge, and $T = 0$ on the straight edge.

It is much better to re-express the problem in terms of the polar coordinates r, θ, because then the boundary conditions will be on constant-coordinate curves: $T = 100$ on $r = c$, $T = 0$ on $\theta = -\pi/2$, and $T = 0$ on $\theta = \pi/2$. Thus, let us re-express (52) in terms of r and θ. First, consider the T_{xx} term. Since $T_{xx} = (\partial/\partial x)(\partial/\partial x)T$, we need to express the differential operator $\partial/\partial x$ in terms of r and θ. By chain differentiation,

$$\frac{\partial(\;)}{\partial x} = \frac{\partial(\;)}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial(\;)}{\partial \theta}\frac{\partial \theta}{\partial x}. \tag{53}$$

The reverse derivatives $r_x, \theta_x, r_y, \theta_y$ are given in (41) and (42) so (53) becomes

$$\frac{\partial}{\partial x} = c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}, \tag{54}$$

where we use the convenient shorthand $c \equiv \cos\theta$ and $s \equiv \sin\theta$. Thus,

$$T_{xx} = \left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)\left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)T$$

$$\begin{aligned} &= \left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)\left(cT_r - \frac{s}{r}T_\theta\right) \\ &= c\frac{\partial}{\partial r}(cT_r) - c\frac{\partial}{\partial r}\left(\frac{s}{r}T_\theta\right) - \frac{s}{r}\frac{\partial}{\partial \theta}(cT_r) + \frac{s}{r}\frac{\partial}{\partial \theta}\left(\frac{s}{r}T_\theta\right) \\ &= c^2 T_{rr} + \frac{cs}{r^2}T_\theta - \frac{cs}{r}T_{\theta r} + \frac{s^2}{r}T_r - \frac{sc}{r}T_{r\theta} + \frac{sc}{r^2}T_\theta + \frac{s^2}{r^2}T_{\theta\theta}. \end{aligned} \tag{55}$$

Similarly,

$$\frac{\partial(\;)}{\partial y} = \frac{\partial(\;)}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial(\;)}{\partial \theta}\frac{\partial \theta}{\partial y} = s\frac{\partial}{\partial r} + \frac{c}{r}\frac{\partial}{\partial \theta}, \tag{56}$$

and (Exercise 13)

$$T_{yy} = s^2 T_{rr} - \frac{cs}{r^2}T_\theta + \frac{cs}{r}T_{\theta r} + \frac{c^2}{r}T_r + \frac{cs}{r}T_{r\theta} - \frac{cs}{r^2}T_\theta + \frac{c^2}{r^2}T_{\theta\theta}. \tag{57}$$

Finally, adding (55) and (57) and recalling that $c^2 + s^2 = 1$ gives

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2} = 0 \tag{58}$$

as the Laplace equation in polar coordinates.

For notational simplicity we used the same letter, T, for the temperature, independent of whether it was regarded as a function of x, y or as a function of r, θ. Strictly speaking, of course, we should have used different letters [as discussed below equation (8) in Section 13.4]: $T(x, y) = T\left(x(r,\theta), y(r,\theta)\right) \equiv U(r,\theta)$, for instance.

Closure. We begin this section by considering a relation $f(x, y) = 0$ on x and y, to determine whether it implies a function $y(x)$ over some neighborhood of x_0, through a point (x_0, y_0) that satisfies the relation. Typically, we cannot solve $f(x, y) = 0$ for y by algebra, so we give up on the idea of a closed form expression of $y(x)$ and seek $y(x)$ in the form of a Taylor series about x_0. We find that all goes well if $f_y(x_0, y_0)$ does not vanish. With that motivational introduction completed, we then state the Implicit Function Theorem 13.6.1, which includes the key stipulation that

$$f_y(x_0, y_0) \neq 0. \tag{59}$$

Generalizing from one relation on two variables to two relations on four variables,

$$\begin{aligned} f(x, y, u, v) &= 0, \\ g(x, y, u, v) &= 0, \end{aligned} \tag{60}$$

we find that the condition (59) generalizes [from the one-by-one determinant on the left-hand side of (59)] to the determinant condition

$$\left.\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}\right|_{x_0, y_0, u_0, v_0} \neq 0,$$

that is, the nonvanishing of the Jacobian determinant of f and g with respect to u and v at (x_0, y_0, u_0, v_0). Similarly for n relations on $2n$ variables, which case is covered in Theorem 13.6.2.

The Jacobians themselves are of great importance. We see that the "reverse derivatives," such as u_y, where u and v are defined implicitly as functions of x and y by $x = x(u, v)$ and $y = y(u, v)$, are expressible as ratios of Jacobian determinants. Further, we see that the result

$$\frac{dy}{dx}\frac{dx}{dy} = 1 \quad \text{or} \quad \frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1},$$

for functions $y(x)$ and $x(y)$, generalizes *not* to $\dfrac{\partial x}{\partial u} = \left(\dfrac{\partial u}{\partial x}\right)^{-1}$ but to the *Jacobian* statements

$$\frac{\partial(x,y)}{\partial(u,v)}\frac{\partial(u,v)}{\partial(x,y)} = 1 \quad \text{or} \quad \frac{\partial(u,v)}{\partial(x,y)} = \left(\frac{\partial(x,y)}{\partial(u,v)}\right)^{-1}, \tag{61}$$

$$\frac{\partial(x,y,z)}{\partial(u,v,w)}\frac{\partial(u,v,w)}{\partial(x,y,z)} = 1 \quad \text{or} \quad \frac{\partial(u,v,w)}{\partial(x,y,z)} = \left(\frac{\partial(x,y,z)}{\partial(u,v,w)}\right)^{-1}, \tag{62}$$

and so on. In fact, in Chapter 15 we will meet Jacobians again, in deriving expressions for area and volume elements for surface and volume integrals.

EXERCISES 13.6

1. Given $f(x, y) = 0$ and a point (x_0, y_0) such that $f(x_0, y_0) = 0$, see if the conditions of Theorem 13.6.1 are met. If so, develop the implicit function $y(x)$ in a Taylor series about x_0, through second-order terms, as we did in Example 4.

(a) $x^2 + y^2 - 2y = 0$; $(1, 1)$
(b) $x^2 + 4y^2 - 4 = 0$; $(0, 1)$
(c) $x^2 + 4y^2 - 4 = 0$; $(0, -1)$
(d) $x(\cos \pi y + 1) + x^3 y + 8 = 0$; $(-2, 1)$
(e) $x(\cos \pi y + 1) + (x^3 + 8)y = 0$; $(-2, 1)$
(f) $x - y - \sin y = 0$; $(0, 0)$
(g) $x - y + \sin y = 0$; $(0, 0)$
(h) $(y - 1)e^y - x^2 + 1 = 0$; $(1, 1)$

2. In each case, find $y'(x)$ and $y''(x)$.

(a) $xy - y^3 = 1$
(b) $xe^y + y = 3$
(c) $y + y^2 + y^3 = x$
(d) $xye^y = 1$
(e) $xy + \sin y = 3x^2$
(f) $y \cos y = x^3$

3. Solve for $y_x|_z$ (i.e., holding z fixed) and $z_x|_y$.

(a) $xy + \sin(x + z) - z^2 = 5$
(b) $xe^y - y^2 - z^2 \sin z = 0$
(c) $e^x + e^y + e^z = 3$
(d) $xy^3 - 2x^3 z + y^5 z^5 = 1$

4. Apply Theorem 13.6.2 to see if implicit functions $u(x, y)$ and $v(x, y)$ exist in some neighborhood of (x_0, y_0), given the values (x_0, y_0, u_0, v_0) satisfying the two relations.

(a) $x - u \cos v = 0$, $y - u \sin v = 0$; $(0, 0, 0, 0)$
(b) $x - u \cos v = 0$, $y - u \sin v = 0$; $(0, 2, 2, \pi/2)$
(c) $x \sin u + y^2 - v^2 = 0$,
$(x + v)^2 - \sin(uy) = 0$; $(1, 1, 0, -1)$
(d) $x \cos u + y + v^2 = 0$, $x - y + \sin(u^2 v) = 0$; $(1, 1, \pi, 0)$
(e) $y \cos u + x - v^3 = 0$, $v^2 + \sin(x - y) = 0$; $(1, 1, \pi, 0)$
(f) $xe^y - 2u^2 + v + 7 = 0$, $\sin y - x^2 \sin v = 0$, $(1, 0, 2, 0)$

5. Evaluate u_y.

(a) $x - y + u^2 + v^2 = 1$, $x + y + u^3 e^v = 2$
(b) $x - u^2 \cos v = 0$, $y - ue^v = 6$

(c) $xe^u - uy + v = 0, \quad x^2u - y^3 + v^3 = 1$
(d) $x - u^2 \sin v = 0, \quad y - u^2 \cos v = 1$

6. Evaluate the indicated Jacobian(s).

(a) $f(u,v) = 3uv^2, \quad g(u,v) = u^2 - v^2; \quad \dfrac{\partial(f,g)}{\partial(u,v)}$
(b) $f(u,v,w) = uw^3, \quad g(u,v,w) = 2v - w,$
$h(u,v,w) = e^{uv}; \quad \dfrac{\partial(f,g,h)}{\partial(u,v,w)}$
(c) $F(p,q) = p^2 + 2, \quad G(p,q) = p \sin q;$
$\dfrac{\partial(F,G)}{\partial(p,q)}, \quad \dfrac{\partial(G,F)}{\partial(p,q)}, \quad \dfrac{\partial(G,F)}{\partial(q,p)}$
(d) $P(x,y) = y^3, \quad Q(x,y) = x^2 - y^2;$
$\dfrac{\partial(P,Q)}{\partial(x,y)}, \quad \dfrac{\partial(P,Q)}{\partial(y,x)}$
(e) $F(x,y,z) = x + y + z, \quad G(x,y,z) = x^2 + y^2 + z^2,$
$H(x,y,z) = x^3 + y^3 + z^3; \quad \dfrac{\partial(F,G,H)}{\partial(x,y,z)}$

7. Show that

$$\text{(a)}\ \frac{\partial(f,g)}{\partial(x,y)} = -\frac{\partial(f,g)}{\partial(y,x)} \qquad \text{(b)}\ \frac{\partial(f,g)}{\partial(x,y)} = \frac{\partial(g,f)}{\partial(y,x)}$$

8. Verify (44), given the relation

(a) $xe^u - x^3u - 5 = 0$
(b) $xu^2 - ue^{-x} - u^5 = 0$
(c) $x + u + x^3 + u^3 - 9 = 0$
(d) $e^{xu} - \sin(x + u) - 1 = 0$

9. Verify (48) for these cases, by working out the eight partial derivatives, multiplying the determinants, and showing that the result is unity.

(a) $x = u + v, \quad y = u - v$
(b) $x = u + v, \quad y = u^3 - v^3$
(c) $x = u + e^v, \quad y = ve^u$
(d) $xe^y - u + e^v = 2, \quad x^2 + y^2 + u^2 + v^2 = 1$
(e) $x + y + u + v = 1, \quad x^2 + y^2 + u^2 + v^2 = 4$

10. (*Chain rule*) Recall from the calculus that if $u = u(x(s))$, then

$$\frac{du}{dx}\frac{dx}{ds} = \frac{du}{ds}, \tag{10.1}$$

which result is an example of the **chain rule**; equation (44) is a special case of (10.1), where s is u. If u and v are functions of x and y, and x and y, in turn, are functions of r and s, then (10.1) generalizes to

$$\boxed{\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(r,s)} = \frac{\partial(u,v)}{\partial(r,s)}.} \tag{10.2}$$

Similarly,

$$\boxed{\frac{\partial(u,v,w)}{\partial(x,y,z)}\frac{\partial(x,y,z)}{\partial(r,s,t)} = \frac{\partial(u,v,w)}{\partial(r,s,t)},} \tag{10.3}$$

and so on. Prove (10.2). HINT: Proceed essentially as we did in (50)-(51).

11. Verify (10.2), above, for these cases.

(a) $x = u\cos v, \quad y = u \sin v,$ and $u = r + s, \quad v = r^2 + s^2$
(b) $x = u + v, \quad y = u - v,$ and $u = r^2 + 2s, \quad v = r\cos s$
(c) $x = u, \quad y = u + v,$ and $u = r\cos s, \quad v = r \sin s$
(d) $x = u^2 + 4v, \quad y = u - v,$ and $u = r - s, \quad v = r^2 + 4s$

12. One speaks of a relation

$$f(p,T,v) = 0 \tag{12.1}$$

on the pressure p, the absolute temperature T, and the specific volume v of a gas as an *equation of state*.

(a) If we think of (12.1) as implicitly defining $v(p,T)$, show that

$$\frac{\partial v}{\partial p} = -\frac{f_p}{f_v} \quad \text{and} \quad \frac{\partial v}{\partial T} = -\frac{f_T}{f_v}, \quad \text{if} \quad f_v \neq 0. \tag{12.2}$$

(b) Is it true, in this case, that

$$\frac{\partial v}{\partial p} = \left(\frac{\partial p}{\partial v}\right)^{-1} \quad \text{and} \quad \frac{\partial v}{\partial T} = \left(\frac{\partial T}{\partial v}\right)^{-1} ? \tag{12.3}$$

Explain.
(c) One well known equation of state is the **van der Waals equation**

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT, \tag{12.4}$$

where a, b, R are constants. For this case, compute v_p and v_T, using (12.2) and, again, using (12.3). Show that the results agree.

13. Derive (57), just as we derived (55).

14. In Section 13.6.4 we show how to transform the Laplace equation (52) under the change of variables $x = r\cos\theta$, $y = r\sin\theta$. Do likewise, for the given change of variables.

(a) $x = 2u + v, \quad y = 2u - v$
(b) $x = 2u + v, \quad y = 3u + v$
(c) $x = e^v, \quad y = u - v^2$

(d) $x = u^2 + v^2, \quad y = u + v$
(e) $x = u^2 \cos v, \quad y = u^2 \sin v$
(f) $x = u^2 + v, \quad y = u^2 - v$

15. In general, if we transform the Laplace equation (52) according to a change of variables

$$x = x(u, v), \quad y = y(u, v),$$

the equation gets changed; i.e., it is *not* of the form

$$T_{uu} + T_{vv} = 0. \tag{15.1}$$

However, show that if the transformation is such that

$$u_x = v_y, \quad u_y = -v_x, \quad \text{and} \quad u_x^2 + u_y^2 \neq 0,$$

then the Laplace equation is *preserved*; i.e., the result *is* (15.1). NOTE: This result is at the heart of the method of **conformal mapping**, to which we devote a later chapter.

13.7 Maxima and Minima

PREREQUISITE: Section 11.6 on quadratic forms is a prerequisite for Section 13.7.2.

13.7.1. Single variable case. Although our interest in this section is in functions of more than one variable, let us begin by reviewing from the calculus the theory of maxima and minima of functions of a single real variable.

We say that a function f has an **absolute maximum** at x_0 if $f(x) \leq f(x_0)$ for all x in the domain of definition of f, and an **absolute minimum** at x_0 if $f(x) \geq f(x_0)$ for all x in the domain. Further, it has a **local maximum** at a point x_0 in its domain if $f(x) \leq f(x_0)$ for all x in some neighborhood of x_0, and a **local minimum** at x_0 if $f(x) \geq f(x_0$ for all x in some neighborhood of x_0.* We use the term **extremum** to mean either a maximum or a minimum. For instance, a local extremum is either a local maximum or a local minimum.

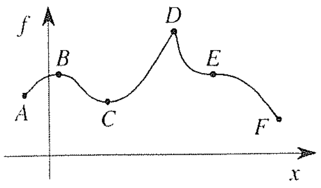


Figure 1. Maxima and minima.

To illustrate, consider the function f, the graph of which is shown in Fig. 1. We can see that f has extrema at A, B,C, D, and F (but not at E): a local minimum at A, a local maximum at B, a local minimum at C, a local and absolute maximum at D, and a local and absolute minimum at F. In this section we will study extrema such as B and C, at which the derivative of f exists and is zero.

THEOREM 13.7.1 *Vanishing Derivative, for Local Extremum*
For a function f of a single real variable x to have a local extremum at a point X in its domain of definition, where f is differentiable at X, it is necessary that

$$f'(X) = 0. \tag{1}$$

Proof: Suppose f has a local maximum at X. Then $f(x) \leq f(X)$ for all x's within some neighborhood N of X. Since f is differentiable at X,

*The term *relative* is sometimes used in place of *local*.

$$\lim_{x \to X} \frac{f(x) - f(X)}{x - X} \tag{2}$$

exists, independent of the manner in which x approaches X. Since $x \to X$, we can assume that x is within N, so $f(x) \leq f(X)$. If $x \to X$ from the right, then

$$\lim_{x \to X+} \frac{f(x) - f(X)}{x - X} = f'(X) \leq 0 \tag{3}$$

because $f(x) - f(X) \leq 0$ and $x - X > 0$, and if $x \to X$ from the left, then

$$\lim_{x \to X-} \frac{f(x) - f(X)}{x - X} = f'(X) \geq 0 \tag{4}$$

because $f(x) - f(X) \leq 0$ and $x - X < 0$. Comparison of (3) and (4) reveals that $f'(X) = 0$. A similar argument applies if f has a local minimum at X. ■

That the condition (1) is not also sufficient is evident from Fig. 1, for $f'(x) = 0$ at E, yet f has neither a local maximum at E nor a local minimum, it has a **horizontal inflection point** there. If the derivative $f'(x)$ is found to vanish at a point X but we do not have the graph of f to examine, how can we determine whether f has a local maximum, local minimum, or horizontal inflection point there?

THEOREM 13.7.2 *Maximum, Minimum, Horizontal Inflection Point*
Suppose that

$$f'(X) = f''(X) = \cdots = f^{(n-1)}(X) = 0,$$

but $f^{(n)}(X) \neq 0$, and that $f^{(n)}(x)$ is continuous in some neighborhood of X, where $n \geq 2$. If n is even, then f has a local maximum at X if $f^{(n)}(X) < 0$ and a local minimum at X if $f^{(n)}(X) > 0$. If n is odd, then f has a horizontal inflection point at X.

Proof: Since by assumption $f^{(n)}(x)$ is continuous in some neighborhood of X, there must exist a neighborhood $N(X)$ throughout which $f^{(n)}(x)$ has the same sign as $f^{(n)}(X)$.* By Taylor's formula with Lagrange remainder, we can express

$$f(x) = f(X) + 0 + \cdots + 0 + \frac{f^{(n)}(\xi)}{n!}(x - X)^n,$$

*If $F(x)$ is continuous at X, then, according to the definition of continuity, to each number $\epsilon > 0$ there corresponds a number $\delta > 0$ such that $|F(x) - F(X)| < \epsilon$ whenever $|x - X| < \delta$. Thus, if we choose ϵ to be smaller than $|F(X)|$, then there exists a $\delta > 0$ such that $F(x)$ has the same sign as $F(X)$ for all x in the neighborhood $|x - X| < \delta$.

where ξ is some point between X and x. If we require that x be in $N(X)$, then ξ must be in $N(X)$ as well. Thus, knowing that $f^{(n)}(\xi)$ has the same sign as $f^{(n)}(X)$ and that

$$f(x) - f(X) = \frac{f^{(n)}(\xi)}{n!}(x - X)^n, \tag{5}$$

then we can see from (5) that:

(1) if n is even and $f^{(n)}(X) > 0$, then $f(x) - f(X) \geq 0$ in $N(X)$, so f has a local minimum at X;

(2) if n is even and $f^{(n)}(X) < 0$, then $f(x) - f(X) \leq 0$ in $N(X)$, so f has a local maximum at X;

(3) and if n is odd, then [independent of the sign of $f^{(n)}(X)$] $f(x) - f(X)$ is positive to one side of X and negative to the other, so f has a horizontal inflection point at X,

which step completes the proof. ■

NOTE: It is sometimes stated that if $f'(X) = 0$, then f has a local maximum, local minimum, or a horizontal inflection point at X if $f''(X)$ is negative, positive, or zero, respectively. No, if $f''(X) = 0$ then no conclusion can be drawn, according to Theorem 13.7.2, until we determine the first nonvanishing derivative, $f^{(n)}(X)$.

EXAMPLE 1. Let $f(x) = (2\sqrt{x} - x - 1)^2$ over $0 < x < \infty$. Then

$$f'(x) = 2\left(\frac{1}{\sqrt{x}} - 1\right)(2\sqrt{x} - x - 1) = -\frac{2}{\sqrt{x}}(1 - \sqrt{x})^3, \tag{6}$$

which vanishes at $x = 1$. Differentiating further, at $x = 1$,

$$f'(1) = 0, \quad f''(1) = 0, \quad f'''(1) = 0, \quad f''''(1) = \frac{3}{2}.$$

Thus, $n = 4$, which is even, and $f''''(1) > 0$, so f has a minimum at $x = 1$. ■

EXAMPLE 2. Let $f(x) = (x - 1)^2 \ln x$ over $0 < x < \infty$. Again $f'(1) = f''(1) = 0$, but $f'''(1) = 6$. This time n is odd so f has a horizontal inflection point at $x = 1$. ■

13.7.2. Multivariable case. Let us extend these ideas to functions $f(x_1, \ldots, x_n)$ of n independent real variables, $x_1, \ldots, x_n$. We continue to be interested in *local* extrema at *interior* points of the domain of f, and we adopt the vector notation $\mathbf{x}$ as shorthand for the point $(x_1, \ldots, x_n)$ in the n-dimensional space, and $f(\mathbf{x})$ as shorthand for $f(x_1, \ldots, x_n)$. We say that f has a **local maximum** at a point $\mathbf{X}$ of its domain if $f(\mathbf{x}) \leq f(\mathbf{X})$ for all points $\mathbf{x}$ in some neighborhood of $\mathbf{X}$, and a **local minimum** at $\mathbf{X}$ if $f(\mathbf{x}) \geq f(\mathbf{X})$ for all points $\mathbf{x}$ in some neighborhood of $\mathbf{X}$.

THEOREM 13.7.3 *Vanishing Partial Derivatives for Local Extremum*
For a function f of n real variables $x_1, \dots, x_n$ to have a local extremum at a point $\mathbf{X} = (X_1, \dots, X_n)$ in its domain of definition, where f is C^1 in some neighborhood of $\mathbf{X}$, it is necessary that

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \ \dots \ , \quad \frac{\partial f}{\partial x_n} = 0 \tag{7}$$

at $\mathbf{X}$.

Proof: Let $x_1 = x_1(\tau), \dots, x_n = x_n(\tau)$ be parametric equations of any curve through $\mathbf{x} = \mathbf{X}$ in n-space, with $\mathbf{x} = \mathbf{X}$ at $\tau = 0$, say. Further, let $F(\tau) \equiv f(\mathbf{x}(\tau))$. Since f has been assumed to be C^1, we can use the chain rule to express

$$\frac{dF}{d\tau} = \frac{\partial f}{\partial x_1}\frac{dx_1}{d\tau} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{d\tau}.$$

Since $f(\mathbf{x})$ has an extremum at $\mathbf{X}$, $F(\tau)$ has an extremum at $\tau = 0$ so, according to Theorem 13.7.1, it is necessary that

$$\frac{dF}{d\tau}(0) = \frac{\partial f}{\partial x_1}(\mathbf{X})\frac{dx_1}{d\tau}(0) + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{X})\frac{dx_n}{d\tau}(0) = 0. \tag{8}$$

Since (8) must hold for *every* path $\mathbf{x} = \mathbf{x}(\tau)$ through $\mathbf{X}$, it follows from (8) that $\dfrac{\partial f}{\partial x_1}(\mathbf{X}) = 0, \dots, \dfrac{\partial f}{\partial x_n}(\mathbf{X}) = 0$, as was to be shown. That is, since $\mathbf{x}(\tau)$ is arbitrary we can choose it so that $\dfrac{dx_1}{d\tau}(0)$ is nonzero and the other derivatives, $\dfrac{dx_2}{d\tau}(0), \dots, \dfrac{dx_n}{d\tau}(0)$, are zero, in which case we learn from (8) that $\dfrac{\partial f}{\partial x_1}(\mathbf{X}) = 0$. Similarly, we can let $\dfrac{dx_2}{d\tau}(0)$ be nonzero and the other derivatives be zero, and learn from (8) that $\dfrac{\partial f}{\partial x_2}(\mathbf{X}) = 0$, and so on. ■

Before continuing, there is a point of rigor to address. Namely, why do we bother introducing the space curve $\mathbf{x} = \mathbf{x}(\tau)$ in the preceding proof? Why not use the simpler looking *differential* form of the chain rule

$$df = \frac{\partial f}{\partial x_1}(\mathbf{X})dx_1 + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{X})dx_n = 0 \tag{9}$$

in place of the *derivative* form (8)? We could argue that since $x_1, \dots, x_n$ are independent variables, the increments $dx_1, \dots, dx_n$ are independent as well. Thus, letting dx_1 be nonzero and $dx_2 = \cdots = dx_n = 0$, we learn from (9) that $\frac{\partial f}{\partial x_1}(\mathbf{X}) = 0$, and so on, as before. The problem with that approach is that the differential version (9) treats df, dx_1, ... , dx_n as though they are finite numbers, whereas they are not. Nonetheless, in Section 13.7.3 we will use (9) for simplicity, with the understanding that we can always change to the (impeccable) derivative form if we wish to.

Let us continue. Any point $\mathbf{X}$ at which (7) holds is called a **critical point** of f. As emphasized below Theorem 13.7.1 for the case of functions of a single variable, the condition (7) [like the condition (1) in Theorem 13.7.1] is necessary but not sufficient for the existence of an extremum of f at $\mathbf{X}$. For instance, if

$$f(x_1, x_2) = x_1^2 - x_2^2, \tag{10}$$

then $\partial f/\partial x_1 = 2x_1 = 0$ all along the line $x_1 = 0$ (i.e., the x_2 axis) in the x_1, x_2 plane, and $\partial f/\partial x_2 = -2x_2 = 0$ all along the line $x_2 = 0$, so both are zero at the intersection of those lines, namely, at the origin, $\mathbf{X} = (0, 0)$. However, along the x_1 axis $f(x_1, 0) = x_1^2$ has a "valley," a minimum, whereas along the x_2 axis $f(0, x_2) = -x_2^2$ has a "mountain," a maximum, so the flat spot on the f surface at $(0, 0)$ is neither a local maximum nor a local minimum. Rather, we call it a "saddle" because it resembles an equestrian saddle, a valley one way and a mountain the other (Fig. 2).

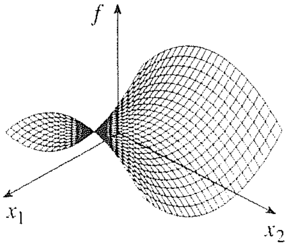


Figure 2. Saddle at $(0, 0)$.

Extending this idea to higher dimensions, we say that $f(\mathbf{x})$ has a **saddle** at a point $\mathbf{X}$ in n-dimensional space if f is C^1 in some neighborhood of $\mathbf{X}$, if $\partial f/\partial x_1 = \cdots = \partial f/\partial x_n = 0$ at $\mathbf{X}$, and if $f(\mathbf{x}) - f(\mathbf{X})$ takes on both positive and negative values in every neighborhood (no matter how small) of $\mathbf{X}$. Of course, if $n = 1$ the saddle is actually a horizontal inflection point, and if $n \geq 3$ we cannot display the graph of f as we have done in Fig. 2 since we would need four or more dimensions.

Given a function $f(x_1, \dots, x_n)$, suppose that, using (7), we determine a critical point $\mathbf{X}$ of f. How can we determine whether f has a local maximum, a local minimum, or a saddle there? Let us explain the essential idea briefly and heuristically.

For a function of a single variable, $f(x)$, with $f'(X) = 0$, we used Taylor's formula to write

$$f(x) = f(X) + \frac{1}{2} f''(\xi)(x - X)^2 \tag{11}$$

for some ξ between X and x so

$$f(x) - f(X) \sim \frac{1}{2} f''(X)(x - X)^2 \tag{12}$$

as $x \to X$. If $f''(X) < 0$ then the right-hand side of (12) represents a "mountain" and f has a local maximum at X; if $f''(X) > 0$ then the right side represents a "valley" and f has a local minimum; if $f''(X) = 0$ no information is obtained and we need to include more terms in Taylor's formula, namely, the first nonvanishing term beyond the initial $f(X)$ term.

For functions of more than one variable the idea is the same. Consider two variables, for simplicity. Given $f(x_1, x_2)$, with $\partial f/\partial x_1 = \partial f/\partial x_2 = 0$ at $\mathbf{X} = (X_1, X_2)$, use Taylor's formula to write

$$\begin{aligned} f(x_1, x_2) - f(X_1, X_2) \sim \frac{1}{2!}\ &\left[f_{x_1x_1}(\mathbf{X})(x_1 - X_1)^2 + f_{x_2x_2}(\mathbf{X})(x_2 - X_2)^2\right. \\ &\left. + 2f_{x_1x_2}(\mathbf{X})(x_1 - X_1)(x_2 - X_2)\right]. \end{aligned} \tag{13}$$

Observe that the right side of (13) is a *quadratic form* in $x_1 - X_1$ and $x_2 - X_2$ so it can be reduced to canonical form as was explained in Section 11.6. Thus, (13) can be re-expressed as

$$f(x_1, x_2) - f(X_1, X_2) \sim \lambda_1 \tilde{x}_1^2 + \lambda_2 \tilde{x}_2^2, \tag{14}$$

where λ_1 and λ_2 are the eigenvalues of the matrix

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix} f_{x_1x_1}(\mathbf{X}) & f_{x_1x_2}(\mathbf{X}) \\ f_{x_1x_2}(\mathbf{X}) & f_{x_2x_2}(\mathbf{X}) \end{bmatrix}, \tag{15}$$

and where $\tilde{\mathbf{x}} = [\tilde{x}_1, \tilde{x}_2]^T$ and $\mathbf{x} = [x_1 - X_1, x_2 - X_2]^T$ are related through the modal matrix transformation $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$. Since (14) expresses $f(x_1, x_2) - f(X_1, X_2)$ as a sum of squares [just as (12) expresses $f(x) - f(X)$ as a square] it reveals whether f has a local maximum, local minimum, or saddle at $\mathbf{X}$: if both λ_1 and λ_2 are negative then f has a local maximum there; if both are positive then f has a local minimum; and if one is positive and one is negative then f has a saddle.

In fact, we have the following theorem.

THEOREM 13.7.4 *Maximum, Minimum, Saddle*
For a function f of n variables $x_1, \dots, x_n$, let

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \quad \dots, \quad \frac{\partial f}{\partial x_n} = 0 \tag{16}$$

at a point $\mathbf{X}$ in f's domain of definition, let f be C^2 in some neighborhood of $\mathbf{X}$, and define the $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} f_{x_1x_1}(\mathbf{X}) & f_{x_1x_2}(\mathbf{X}) & \cdots & f_{x_1x_n}(\mathbf{X}) \\ f_{x_2x_1}(\mathbf{X}) & f_{x_2x_2}(\mathbf{X}) & \cdots & f_{x_2x_n}(\mathbf{X}) \\ \vdots & & & \vdots \\ f_{x_nx_1}(\mathbf{X}) & f_{x_nx_2}(\mathbf{X}) & \cdots & f_{x_nx_n}(\mathbf{X}) \end{bmatrix}. \tag{17}$$

Let det$\mathbf{A}$ be nonzero. If $\mathbf{A}$ is positive definite then f has a local minimum at $\mathbf{X}$, if it is negative definite then f has a local maximum at $\mathbf{X}$, and if it has at least one positive eigenvalue and at least one negative eigenvalue then f has a saddle at $\mathbf{X}$.

pelo menos

Proof: Before beginning the proof let us make several remarks. First, the 1/2 present in (15) has been omitted in (17) because the theorem concerns itself only with the *sign* of the eigenvalues, not with their magnitude. Also, observe that $f_{x_j x_k}(\mathbf{X}) = f_{x_k x_j}(\mathbf{X})$ because f has been assumed to be C^2 in some neighborhood of $\mathbf{X}$; thus, $\mathbf{A}$ is symmetric. Finally, why do we ask $\det\mathbf{A}$ to be nonzero? If $\det\mathbf{A} = 0$, then $\lambda = 0$ is among the eigenvalues of $\mathbf{A}$, which borderline case we wish to avoid. For recall from Theorem 13.7.2 (for the single-variable case) that if $f'(X) = 0$ *and* $f''(X) = 0$ then we need to keep going until we come to the first *non*vanishing derivative at X before we can tell if f has a maximum, minimum, or saddle at X. For the more difficult multivariable case we prefer to avoid such additional complications. In any case, $\det\mathbf{A} \neq 0$ is the generic case, so we are not giving up very much.

To prove the theorem we begin with Taylor's formula

$$f(\mathbf{x}) - f(\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{x_i x_j}(\xi)(x_i - X_i)(x_j - X_j) \tag{18}$$

with the first-order terms omitted by virtue of (16), where $\mathbf{x} = (x_1, \dots, x_n)$, $\mathbf{X} = (X_1, \dots, X_n)$, and $\xi = (\xi_1, \dots, \xi_n)$ is some point on the line connecting $\mathbf{x}$ and $\mathbf{X}$.* Since f is C^2 in some neighborhood of $\mathbf{X}$, there is some neighborhood $N(\mathbf{X})$ throughout which the $f_{x_i x_j}(\xi)$ coefficients in (18) have the same sign as $f_{x_i x_j}(\mathbf{X})$. Thus, in place of the right side of (18) we can use the quadratic form

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{x_i x_j}(\mathbf{X})(x_i - X_i)(x_j - X_j) \tag{19}$$

to determine if $f(\mathbf{x}) - f(\mathbf{X})$ is less than or equal to zero, greater than or equal to zero, or of mixed sign in $N(\mathbf{X})$ and, from the theory of quadratic forms, these circumstances correspond to $\mathbf{A}$ being negative definite, positive definite, or having eigenvalues of mixed sign, respectively. ■

EXAMPLE 3. Identify and classify any local extrema and saddles of the function

$$f(x, y) = \ln\left[2x(y - 1) + 1\right]. \tag{20}$$

(Here, x_1 is x and x_2 is y.) Setting

$$f_x(x, y) = \frac{2(y-1)}{2x(y-1)+1} = 0 \quad \text{and} \quad f_y(x, y) = \frac{2x}{2x(y-1)+1} = 0 \tag{21}$$

gives the single critical point $x = 0$, $y = 1$. Next,

$$f_{xx}(x, y) = -\frac{4(y-1)^2}{[2x(y-1)+1]^2},$$

*That is, $\xi_1 = x_1 + (X_1 - x_1)\tau, \dots, \xi_n = x_n + (X_n - x_n)\tau$ for some τ in $0 \leq \tau \leq 1$. These are parametric equations of a **straight line** from $\mathbf{x}$ to $\mathbf{X}$ in n-space.

$$f_{yy}(x,y) = -\frac{4x^2}{[2x(y-1)+1]^2}, \tag{22}$$
$$f_{xy}(x,y) = \frac{2}{[2x(y-1)+1]^2},$$

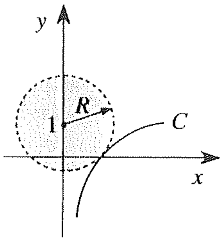


Figure 3. Neighborhood of (0,1).

so we see from (20)–(22) that $f, f_x, f_y, f_{xx}, f_{yy}$, and f_{xy} are continuous everywhere in the x, y plane except along the curve $2x(y-1)+1=0$ (shown as $\mathcal{C}$ in Fig. 3), along which they are undefined. Thus, f is C^2 in any neighborhood, about (0,1), of radius R or less (Fig. 3). Next,

$$\mathbf{A} = \begin{bmatrix} f_{xx}(0,1) & f_{xy}(0,1) \\ f_{xy}(0,1) & f_{yy}(0,1) \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \tag{23}$$

with eigenvalues $\lambda = +2, -2$. The latter are of mixed sign so, according to Theorem 13.7.4, f has a saddle at (0,1).

COMMENT 1. We did not bother verifying that det**A** $\neq 0$ because that condition merely served to disallow zero as an eigenvalue and, indeed, we found that $\lambda = \pm 2$.

COMMENT 2. Even without reducing the quadratic form

$$f(x,y) - f(0,1) \sim \frac{1}{2!}\left[f_{xx}(0,1)x^2 + f_{yy}(0,1)(y-1)^2 + 2f_{xy}(0,1)x(y-1)\right], \tag{24}$$

or,

$$2\left[f(x,y) - f(0,1)\right] \sim \left[0x^2 + 0(y-1)^2 + 4x(y-1)\right] = 4x(y-1), \tag{25}$$

to canonical form, we can see from (25) that $f(x,y) - f(0,1)$ is of mixed sign in every neighborhood of (0,1) as indicated in Fig. 4. ▮

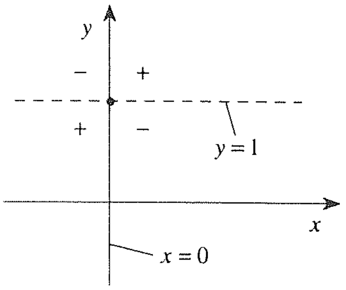


Figure 4. Mixed sign of $2x(y-1)$ near (0,1).

EXAMPLE 4. Consider

$$f(x,y,z) = \sin\left(x^2+y^2+z^2\right) + xy + xz + yz. \tag{26}$$

Working out f_x, f_y, and f_z, it is easily found that (0,0,0) is a critical point of f (although not the only one; see Exercise 9). Working out the second-order partial derivatives as well, we find that f is indeed C^2 in a neighborhood of (0,0,0); in fact, it is C^2 everywhere. Further, $f_{xx} = f_{yy} = f_{zz} = 2$ and $f_{xy} = f_{xz} = f_{yz} = 1$ at (0,0,0) so

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \tag{27}$$

the eigenvalues of which are $\lambda = 1, 1, 4$. Since all of the λ's are positive, f has a local minimum at the critical point (0,0,0). ▮

EXAMPLE 5. Consider

$$f(x,y) = x^3y - xy^3. \tag{28}$$

Then (Exercise 4) there is only one critical point, at (0,0), and $f_{xx} = f_{yy} = f_{xy} = 0$ there, so

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{29}$$

Since det$\mathbf{A} = 0$ (and $\lambda = 0, 0$), Theorem 13.7.4 *does not apply*; (0,0) is said to be a **higher-order critical point**.

Of course, higher-order critical points occur for functions of a single variable as well, as in Example 1, where we had $f''(1) = f'''(1) = 0$ and we had to look to $f''''(1)$ for clarification. But whereas Theorem 13.7.2 covered such higher-order critical points, Theorem 13.7.4 does not.

In the present example inspection happens to suffice, for if we factor f as $xy(x - y)(x + y)$ we can see that f is alternately positive and negative in 45° wedges as shown in Fig. 5. Hence, the critical point of f at (0,0) is a saddle. ▮

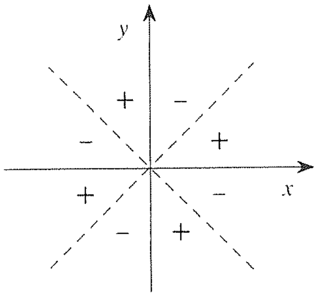


Figure 5. Sign of $f(x, y) = x^3y - xy^3$.

An important physical application of these ideas involves equilibrium states and their stability, because many force fields (such as gravitational force fields) have associated with them a **potential energy** $V(x, y, z)$, such that $\partial V/\partial x$, $\partial V/\partial y$, $\partial V/\partial z$ are the x, y, z force components, respectively, at any field point (x, y, z). Thus, the critical points of V are the **equilibrium points** of the force field, points at which the force is zero. If V has a local minimum at such a point, then the equilibrium at that point is classified as stable; if V has a local maximum or saddle there, then the equilibrium at that point is unstable.

For example, it is known from **Coulomb's law** of electrostatics that the **electric potential** $V(x, y, z)$ at any field point x, y, z, induced by a charge of strength Q coulombs (where Q can be positive or negative) at x_0, y_0, z_0 is proportional to Q and inversely proportional to the distance between x, y, z and x_0, y_0, z_0. Letting the constant of proportionality be k, then

$$V(x, y, z) = k\frac{Q}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}}, \tag{30}$$

and the force components induced at any point x, y, z in the coordinate directions, per unit charge at x, y, z, are given by $F_x = \partial V/\partial x$, $F_y = \partial V/\partial y$, $F_z = \partial V/\partial z$, respectively.

EXAMPLE 6. *Equilibrium Points of Charge Distribution.* Let three charges, each of strength Q coulombs, be placed in the x, y plane at the points (1,0), (0,0), and (0,1), respectively (Fig. 6). Find any equilibrium points and determine their stability. Superimposing the potentials induced by the individual charges gives

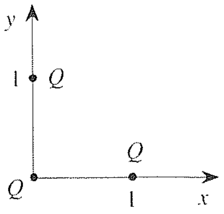


Figure 6. Three charges.

$$V(x, y) = kQ\left(\frac{1}{\sqrt{x^2 + y^2}} + \frac{1}{\sqrt{(x - 1)^2 + y^2}} + \frac{1}{\sqrt{x^2 + (y - 1)^2}}\right). \tag{31}$$

Using computer algebra we can evaluate $\partial V/\partial x$ and $\partial V/\partial y$ and solve the equations

$$\frac{\partial V}{\partial x}(x, y) = 0, \quad \frac{\partial V}{\partial y}(x, y) = 0 \tag{32}$$

for x, y; the *Maple* commands are given at the end of this section. The result is

$$X = 0.0475, \qquad Y = 0.5130. \tag{33}$$

Next, evaluating $\partial^2 V/\partial x^2$, $\partial^2 V/\partial x \partial y$, $\partial^2 V/\partial y^2$ at that point gives, for the $\mathbf{A}$ matrix in (17),

$$\mathbf{A} = \begin{bmatrix} V_{xx}(X,Y) & V_{xy}(X,Y) \\ V_{xy}(X,Y) & V_{yy}(Y,Y) \end{bmatrix} = \begin{bmatrix} -14.375 & -1.448 \\ -1.448 & 31.014 \end{bmatrix}, \tag{34}$$

with eigenvalues

$$\lambda = -14.42, \quad 31.06. \tag{35}$$

Since the λ's are of mixed sign the equilibrium point (33) is unstable. Physically, that means that if a charge is placed at that point, then even the slightest displacement will cause it to move away.

COMMENT. Does it not seem peculiar that the equilibrium point (33) is not on the axis of symmetry, the line $y = x$? This point is discussed at the end of the section when we give the *Maple* commands used in this example. ■

13.7.3. Constrained extrema and Lagrange multipliers. In Section 13.7.2 we studied the problem of finding one or more points, if any, that extremize a given function f:

$$f(x_1, \ldots, x_n) = \text{extremum}, \tag{36}$$

where, by an **extremum** we mean a local maximum or a local minimum. Now, however, we consider the problem of extremizing a given function subject to one or more **constraints**.

To illustrate, suppose we wish to find the point on the plane $2x - y + 4z = 3$ that is closest to the origin. Since the distance from the origin to (x, y, z) is $\sqrt{x^2 + y^2 + z^2}$, we wish to minimize the function $\sqrt{x^2 + y^2 + z^2}$ subject to the constraint that x, y, z are related according to $2x - y + 4z = 3$. Surely, if we minimize the distance we will also be minimizing the square of the distance, $x^2 + y^2 + z^2$, so let us state the problem in the more convenient form

$$f(x, y, z) = x^2 + y^2 + z^2 = \text{minimum}, \tag{37a}$$

subject to the constraint

$$g(x, y, z) = 2x - y + z = 3. \tag{37b}$$

There is only one function being extremized, but there could be more than one constraint:

$$f(x_1, \ldots, x_n) = \text{extremum}, \tag{38a}$$

subject to the constraints

$$\begin{aligned} g_1(x_1, \ldots, x_n) &= c_1, \\ &\vdots \\ g_k(x_1, \ldots, x_n) &= c_k. \end{aligned} \tag{38b}$$

In this discussion let $n = 3$, for definiteness, and let $k = 1$; other cases will be found in the exercises, and the choice $n = 3$ and $k = 1$ should suffice in explaining the main ideas. Thus, we consider the problem

$$f(x, y, z) = \text{extremum}, \tag{39a}$$

$$g(x, y, z) = c, \tag{39b}$$

where we are using the simpler notation x, y, z in place of x_1, x_2, x_3. We assume, throughout this Section 13.7.3 that f and g are C^1 in some neighborhood of the point under consideration, at which f has an extremum.

To begin, we write

$$df = \frac{\partial f}{\partial x}(\mathbf{X})dx + \frac{\partial f}{\partial y}(\mathbf{X})dy + \frac{\partial f}{\partial z}(\mathbf{X})dz = 0 \tag{40}$$

at $\mathbf{X}$. [That is, let us use the more convenient differential form of the chain rule, as in (9), rather than the derivative form, as in (8).] However, and this is a key point, we cannot use the arbitrariness of the dx, dy, dz increments to infer that $\partial f/\partial x = \partial f/\partial y = \partial f/\partial z = 0$ at X, because those increments are *not* arbitrary. Rather, x, y, z are related through the constraint equation (39b) so the dx, dy, dz increments are related as well.

One way out of this difficulty is to solve (39b) for z as a function of x and y, which will be possible if $g_z(\mathbf{X}) \neq 0$. Then, if (39b) gives us $z = Z(x, y)$, we can put that result into (39a) and write

$$f(x, y, Z(x, y)) \equiv F(x, y) = \text{extremum}. \tag{41}$$

Although it is true that x, y, z are not independent variables (for they are related through the constraint equation), x and y, by themselves, *are* independent.* Thus,

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = 0 \tag{42}$$

does imply that

$$\frac{\partial F}{\partial x} = 0, \tag{43a}$$

$$\frac{\partial F}{\partial y} = 0, \tag{43b}$$

because dx and dy in (42) are arbitrary independent increments. The idea, then, is to solve (43a,b) for x and y; then, $z = Z(x, y)$ gives z.

EXAMPLE 7. Let us use this method to solve the problem stated in (37), above. Solving (37b) for z gives $z = 3 - 2x + y$ so

$$f(x, y, Z(x, y)) = x^2 + y^2 + (3 - 2x + y)^2 \equiv F(x, y). \tag{44}$$

*That is, we can assign values to x and y independently. Then, z is not arbitrary but follows from the constraint equation (39b).

Then,

$$\begin{aligned} \frac{\partial F}{\partial x} &= 2x - 4(3 - 2x + y) = 0, \\ \frac{\partial F}{\partial y} &= 2y + 2(3 - 2x + y) = 0 \end{aligned} \tag{45}$$

so $x = 1$ and $y = -1/2$. Finally, $z = 3 - 2x + y = 3 - 2 - 1/2 = 1/2$. To verify that $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ does give a minimum [as specified in (37a)], and not a maximum or a saddle, we can apply Theorem 13.7.4 to $F(x, y)$. We find that

$$\mathbf{A} = \begin{bmatrix} F_{xx} & F_{xy} \\ F_{xy} & F_{yy} \end{bmatrix} = \begin{bmatrix} 10 & -4 \\ -4 & 4 \end{bmatrix}$$

has eigenvalues $\lambda = 2$ and 12. Since both are positive, $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ does give a minimum, as desired. ▮

The essential idea behind the method described above is to use the constraint equations to eliminate variables so that those remaining are independent. Let us call that procedure the **method of elimination**, to distinguish it from the **method of Lagrange multipliers** due the the great French mathematician *Joseph Louis Lagrange* (1736–1813). Like Euler, Lagrange also worked on the applications of mathematics to mechanics.

To explain Lagrange's method, consider once again the representative problem (39). Besides df vanishing at an extremum, so does dg (indeed, $dg = 0$ everywhere because g is a constant). Thus,

$$df = f_x dx + f_y dy + f_z dz = 0, \tag{46a}$$

$$dg = g_x dx + g_y dy + g_z dz = 0, \tag{46b}$$

from which it follows that

$$df - \lambda dg = (f_x - \lambda g_x)dx + (f_y - \lambda g_y)dy + (f_z - \lambda g_z)dz = 0, \tag{47}$$

where λ is a yet to be determined parameter, the so-called **Lagrange multiplier**. Remember that x, y, z are not independent variables because they are related through (39b). Thus, we cannot argue that the dx, dy, dz increments in (47) are arbitrary and hence that each of the coefficients $f_x - \lambda g_x$, $f_y - \lambda g_y$, $f_z - \lambda g_z$ must vanish at the extremum $\mathbf{X}$. However, we can reach that same conclusion by a different logic. For suppose that $g_z \neq 0$ at $\mathbf{X}$. Then, by choosing λ to be f_z/g_z (which number exists because $g_z \neq 0$ at $\mathbf{X}$), (47) reduces to

$$(f_x - \lambda g_x)dx + (f_y - \lambda g_y)dy = 0. \tag{48}$$

[When we write $\lambda = f_z/g_z$, we do not fear that we are setting a constant equal to a function of x, y, z, for (46) and (47) hold *at* $\mathbf{X}$, so f_x, g_x, f_y, ..., really mean $f_x(\mathbf{X})$, $g_x(\mathbf{X})$, $f_y(\mathbf{X})$, ... ; that is, they are numbers, not functions.] Now, since x and y, by themselves, are independent variables, we can argue that dx and dy

in (48) *are* arbitrary independent increments so that the coefficients $f_x - \lambda g_x$ and $f_y - \lambda g_y$ must vanish at $\mathbf{X}$.

Consequently, we have the four equations

$$f_x - \lambda g_x = 0, \qquad f_y - \lambda g_y = 0, \qquad f_z - \lambda g_z = 0 \tag{49a}$$

and

$$g = c \tag{49b}$$

on the four unknowns x, y, z, and λ. In effect, the Lagrange multiplier method amounts to forming a new function,

$$\boxed{f^* \equiv f - \lambda g} \tag{50}$$

(or $f + \lambda g$; it doesn't matter), and extremizing f^* subject to *no* constraints so that $f_x^* = f_y^* = f_z^* = 0.$, which equations are identical to (49a). Observe that even though we already used the constraint equation $g = c$ in obtaining (46b), we still need to append $g = c$ to equations (49) because we lost information when we took the differential of $g = c$, in (46b), in that the constant c thereby dropped out.

EXAMPLE 7. Let us solve the problem stated in (37) once again, this time using the Lagrange multiplier method. With

$$f^* = f - \lambda g = x^2 + y^2 + z^2 - \lambda(2x - y + z),$$

we have

$$\begin{aligned} f_x^* &= 2x - 2\lambda = 0, \\ f_y^* &= 2y + \lambda = 0, \\ f_z^* &= 2z - \lambda = 0 \end{aligned} \tag{51a}$$

and

$$g = 2x - y + z = 3. \tag{51b}$$

Solving (51) gives $x = 1, \; y = -1/2, \; z = 1/2$, and $\lambda = 1$. Thus, $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ as found in Example 7 by the method of elimination.

COMMENT. In this problem we have no special interest in the ultimate value of λ; λ is simply an auxiliary variable. However, in physical applications λ often turns out to have physical significance. For instance, in studying the dynamics of the motion of a bead along a curved frictionless wire λ might turn out to be the normal force between the bead and the wire. ▮

EXAMPLE 8. Suppose that we wish to construct a tin can in the shape of a rectangular prism. Let the base dimensions be x by y and let the height be z, such that the volume xyz is a prescribed value V. The base is to be twice as thick as the other five sides so its cost per unit area is 2α, where α is the cost per unit area of the other sides. Seeking to minimize the material cost gives the problem:

$$\text{Material cost:} \quad f(x, y, z) = \alpha(3xy + 2xz + 2yz) = \text{minimum}, \tag{52a}$$

subject to the constraint

$$\text{Volume:} \quad g(x,y,z) = xyz = V. \tag{52b}$$

Defining $f^* = f - \lambda g$, set

$$\begin{aligned} f_x^* &= f_x - \lambda g_x = \alpha(3y+2z) - \lambda yz = 0, \\ f_y^* &= f_y - \lambda g_y = \alpha(3x+2z) - \lambda xz = 0, \\ f_z^* &= f_z - \lambda g_z = \alpha(2x+2y) - \lambda xy = 0. \end{aligned} \tag{53}$$

Solving the four equations (52b) and (53) gives $x = y = (2V/3)^{1/3}$, $z = \frac{3}{2}(2V/3)^{1/3}$. ∎

Closure. In this section we review the theory of maxima and minima for functions of a single variable before turning to the unconstrained and constrained extremization of functions of more than one variable, considering only local extrema at interior points of the domain of the function f being extremized. If $f(x_1,\dots,x_n)$ is C^1 in some neighborhood of the point of extremum, $\mathbf{X}$, then the extremum condition $df = 0$ implies that we must have

$$\frac{\partial f}{\partial x_1} = \cdots = \frac{\partial f}{\partial x_n} = 0 \tag{54}$$

at $\mathbf{X}$. Points at which (54) holds are called critical points of f. However, f need not have an extremum (local maximum or local minimum) at a critical point; it could have a saddle there. Assuming a bit more of f (namely, that it be C^2 rather than C^1 in some neighborhood of $\mathbf{X}$) and looking to the second-order terms in Taylor's formula, we show in Theorem 13.7.4 that everything hinges on the eigenvalues of the matrix $\mathbf{A} = \{f_{x_i x_j}(\mathbf{X})\}$ associated with f: if all of $\mathbf{A}$'s eigenvalues are negative then f has a local maximum at $\mathbf{X}$, if all are positive then f has a local minimum there, and if they are mixed in sign then f has a saddle there.

For the extremization of f in the presence of one or more constraints of the form $g = c$, the extremum condition $df = 0$ still holds but no longer implies (54). We study two methods of solving such problems, the method of elimination and the method of Lagrange multipliers. Working with three independent variables (x, y, z) for definiteness, we find that both methods work if $g_z(\mathbf{X}) \neq 0$ [or, failing that, if either $g_y(\mathbf{X}) \neq 0$ or $g_z(\mathbf{X}) \neq 0$].

In closing, let us mention an important type of extremization problem that does not fit the categories covered in this section. Namely, we seek $x_1, \dots, x_n$ so that

$$f(x_1,\dots,x_n) = b_1x_1 + \cdots + b_nx_n = \text{extremum}, \tag{55a}$$

subject to the constraints

$$\begin{aligned} g_1(x_1,\dots,x_n) &= a_{11}x_1 + \cdots + a_{1n}x_n = c_1, \\ &\vdots \\ g_k(x_1,\dots,x_n) &= a_{k1}x_1 + \cdots + a_{kn}x_n = c_k, \end{aligned} \tag{55b}$$

where all the x_j's are required to be *nonnegative*. Or the constraints might be *inequalities* such as

$$a_{11}x_1 + \cdots + a_{1n}x_n \leq c_1, \tag{56}$$

but this case is equivalent to the one stated above since we can introduce an additional variable x_{n+1} and rewrite (56) as

$$a_{11}x_1 + \cdots + a_{1n}x_n + x_{n+1} = c_1$$

with the proviso that x_{n+1} is nonnegative too; x_{n+1} is called a **slack variable**.

Such problems are of great interest in the field known as **operations research**, often abbreviated as **OR**. Solutions are obtained not by setting various partial derivatives equal to zero but by special techniques such as the **simplex method**, which fall under the general heading of **linear programming**. This topic is beyond our present scope.*

Computer software. Let us go through the *Maple* commands used in Example 6. First, define the function V by entering

$$V := (x\text{^}2+y\text{^}2)\text{^}(-1/2)+((x-1)\text{^}2+y\text{^}2)\text{^}(-1/2)+(x\text{^}2+(y-1)\text{^}2)\text{^}(-1/2)) :$$

We can obtain $\partial V/\partial x$ by a diff(V, x) command, but since we will subsequently wish to call that quantity, let us both evaluate it and give it the name Vx at the same time. Similarly for $\partial V/\partial y$. Thus, enter

$$Vx := \text{diff}(V, x) :$$

and

$$Vy := \text{diff}(V, y) :$$

(Whereas we've been using colons you may prefer semicolons so that the results will be printed.) To find the solution of (32), use the fsolve command, and define the solution(s) as X, Y at the same time:

$$XY := \text{fsolve}(\{Vx = 0, Vy = 0\}, \{x, y\}, \{x = 0..1, y = 0..1\});$$

where we have used the optional $\{x = 0..1, y = 0..1\}$ to narrow the solution search because it can be seen by inspection (Exercise 11) that the equilibrium point(s), if any, must lie in that unit square. The output is

$$XY := \{y = .5129880394, x = .04746059840\}$$

Next,

$$Vxx : = \text{diff}(V, x, x);$$
$$Vxy : = \text{diff}(V, x, y);$$
$$Vyy : = \text{diff}(V, y, y);$$

*See, for instance, G. B. Dantzig, *Linear Programming and Extensions*, 2nd ed. (Princeton, NJ: Princeton University Press, 1963) or R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (New York: Wiley, 1988).

give $\partial^2 V/\partial x^2$, $\partial^2 V/\partial x \partial y$, $\partial^2 V/\partial y^2$. To evaluate these second-order derivatives at X, Y and define them as the elements of the $\mathbf{A}$ matrix, enter

$$\begin{aligned} a11 &:= \text{eval(subs}(XY, Vxx));\\ a12 &:= \text{eval(subs}(XY, Vxy));\\ a22 &:= \text{eval(subs}(XY, Vyy)); \end{aligned}$$

which gives

$$\begin{aligned} a11 &:= -14.37506987\\ a12 &:= -1.448243172\\ a22 &:= 31.01365215 \end{aligned}$$

respectively. Then enter

with(linalg):

to access the linear algebra package. Finally,

$$A := \text{matrix}(2, 2, [a11, a12, a12, a22]);$$

and

eigenvals(”);

give the eigenvalues

$$-14.42123281, \quad 31.05981510$$

as stated in the example.

Note the use of the "**subs**" in the eval commands. In case that usage is not clear, let us give another example: to evaluate $3x^2$ at $x = 5$ use the command eval(subs(5,3 $*$ x^2)). Note also the strange outcome, that the equilibrium point is not on the line of symmetry $y = x$ even though both the potential $V(x, y)$ is a symmetric function of x and y or, said differently, even though the physical arrangement of charges is symmetric about that line. (You might want to think about that paradox before reading on.) However, we must realize that the fsolve command does not necessarily give all solutions; in fact, if we do not use the option to narrow the search to the unit square then the command gives no solutions at all. If there *is* symmetry about the line $y = x$ then there must be at least one more solution within the unit square, at the point $x = 0.5129880394, y = 0.04746059840$. To test that idea we can change the $\{x = 0..1, y = 0..1\}$ option to $\{x = 0.5..0.52, y = 0.045..0.050\}$, say. That change does indeed produce the suspected solution, so there are *two* equilibrium points and they are arranged symmetrically about the line $y = x$.

EXERCISES 13.7

1. The given function has a critical point at $x = 1$. Classify it as a local maximum, local minimum, or horizontal inflection point.

(a) $(x-1)^2 \ln x$ (b) $3(x-1)^4 + 5$
(c) $x(x-2)(x-1)^2$ (d) $(x+1)(x-3)(1-x)^3$
(e) $\sin[5(x-1)^4]$ (f) $\exp[8(x-1)^5]$
(g) $(1-x)\sin[(x^2-1)^3]$ (h) $\exp[-(\ln x)^3]$

2. Find all critical points of the function and classify them as local maxima, local minima, or horizontal inflection points.

(a) $\exp(-x^3)$, $-\infty < x < \infty$
(b) $1/(x^2 - 4x + 5)$, $-\infty < x < \infty$
(c) $\exp(12x - x^3)$, $-\infty < x < \infty$
(d) $-3(\ln x)^3$, $0 < x < \infty$
(e) $(\ln x)^4$, $0 < x < \infty$
(f) $\exp(-\sin x)$, $-1 < x < 5$
(g) $x^2 e^{-x}$, $-\infty < x < \infty$
(h) $4\sqrt{x} - x^2$, $0 < x < \infty$

3. Show that if $f(x)$ and $g(x)$ have local maxima at $x = X$, then the product $f(x)g(x)$ has a critical point there, which can be a local maximum, a local minimum, or a horizontal inflection point.

4. Show that the only critical point of $f(x,y) = x^3y - xy^3$ is (0,0) as claimed in Example 5.

5. Find, and classify, all local maxima, local minima, and saddles for the given function of x and y. Use Theorem 13.7.4 if possible.

(a) $2x^2 + xy + y^2 + 7y + 8$ (b) $x^2 + xy + y^2 + 6y$
(c) $x^2 + 4xy + y^2 + 6x + 8$ (d) $-6xy + x^5$
(e) $x(2x - y)$ (f) $\exp[-(x^2 + y^2 + 1)]$
(g) $\ln(x^2 + xy + y^2 + 4)$ (h) $\exp(x^2 - y^2)$
(i) $2xy - \sin xy$ (j) $x^3 - xy + x^5$
(k) $\ln(1 + x^2 + y^2)$ (l) $x^3 - xy + x^2y - 2$
(m) $1/(xy + y^2 + y + 1)$ (n) $2xy + x^8$

6. Same as Exercise 5.

(a) $f(x,y,z) = x^2 + 3xy + z^2$
(b) $f(x,y,z) = \exp(2x^2 + xz - 5z^2)$
(c) $f(x,y,z) = x^2 + y^2 + z^2 + xy + xz + yz$
(d) $f(x,y,z) = \ln(x^2 + y^2 + z^2 + 1)$
(e) $f(w,x,y,z) = \exp(w^2 - wx - yz + z^2)$
(f) $f(w,x,y,z) = 7 - 4(w^2 + x^2 + y^2 + z^2) + 4wx + 2wz$

7. Is it true that $f(x,y,z) = 1 + x^2 + 5y^2 + 2(z-1)^2 + 2\sin[x(z-1)]$ has a local minimum at (0,0,1)?

8. Let $f(x,y,z) = (1 + x^2 + y^2)^5 + \alpha \sin(xy) + 6z^2(1-z)$. For what values of α, if any, will f have a local minimum at (0,0,0)?

9. In Example 4 we considered the critical point at (0,0,0). Show that f has infinitely many other critical points as well. Find them and show that they comprise an infinite set of circles in 3-space. HINT: You will need to find nontrivial solutions of the equations

$$\begin{aligned} 2cx + y + z &= 0, \\ x + 2cy + z &= 0, \\ x + y + 2cz &= 0, \end{aligned}$$

where $c \equiv \cos(x^2 + y^2 + z^2)$.

10. For the case of a function of only two variables, $f(x,y)$, one can find the eigenvalues of the $\mathbf{A}$ matrix in (17) by the quadratic formula and examine their signs. Doing so, one can use Theorem 13.7.4 to obtain the following more specialized and frequently stated theorem, which we ask you to prove.

THEOREM 13.7.5 *Maximum, Minimum, Saddle*
Let

$$f_x(a,b) = 0 \quad \text{and} \quad f_y(a,b) = 0, \tag{10.1}$$

where (a,b) is an interior point in the domain of definition of f, and let f be C^2 in some neighborhood of (a,b). Define

$$D \equiv \begin{vmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{xy}(a,b) & f_{yy}(a,b) \end{vmatrix}. \tag{10.2}$$

(i) If $f_{xx}(a,b) > 0$ and $D > 0$, then f has a local minimum at (a,b).
(ii) If $f_{xx}(a,b) < 0$ and $D > 0$, then f has a local maximum at (a,b).
(iii) If $D < 0$, then f has a saddle at (a,b).

11. (*More about Example 6*) (a) In Example 6 we state that it can be seen by inspection that the equilibrium point(s), if any, must lie within the unit square $0 < x < 1$, $0 < y < 1$. Explain why that must be so.
(b) To understand the result obtained in Example 6, draw a

neat approximate sketch of the forces acting on a charge (of the same sign as Q, say) at the equilibrium point $x \approx 0.047, y \approx 0.513$, showing how the forces do indeed balance at that point. Also, explain, in physical terms, why the equilibrium is unstable.
(c) Repeat the calculation of Example 6, again using computer software, with the charge at the origin removed.
(d) Repeat the calculation of Example 6, again using computer software, with another charge Q added at the point $x = 1, y = 1$.
(e) Repeat the calculation of Example 6, again using computer software, but for a charge configuration supplied by your instructor.

12. (*Linear least-squares fit*) Suppose that the values $f(x_1) = f_1, f(x_2) = f_2, \dots, f(x_N) = f_N$ are known; e.g., they may be experimental datum points. And suppose that we wish to find the best linear fit $f(x) \approx ax + b \equiv F(x)$ to these data, best in the sense that the total squared error E is minimized:

$$E(a,b) = \sum_{j=1}^{N} [F(x_j) - f_j]^2 = \text{minimum}. \tag{12.1}$$

It may be assumed that $x_1, \dots, x_N$ are distinct points.
(a) Derive the results

$$a = \frac{N\sum x_j f_j - (\sum f_j)(\sum x_j)}{N\sum x_j^2 - (\sum x_j)^2}, \tag{12.2}$$

$$b = \frac{(\sum x_j^2)(\sum f_j) - (\sum x_j)(\sum x_j f_j)}{N\sum x_j^2 - (\sum x_j)^2}, \tag{12.3}$$

where $\sum$ denotes $\sum_{j=1}^{N}$.
(b) Verify that (12.2) and (12.3) give a local minimum of $E(a,b)$ and also that the denominators in (12.2) and (12.3) cannot vanish. HINT: The Schwarz inequality $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\|\,\|\mathbf{v}\|$, where $\mathbf{u} = (u_1, \dots, u_N)$, $\mathbf{v} = (v_1, \dots, v_N)$ should be useful if $\mathbf{u}$ and $\mathbf{v}$ are suitable chosen. Also, you may use Theorem 13.7.5 in Exercise 10 if you wish.
(c) Show that the local minimum of $E(a,b)$ is, in fact, an *absolute* minimum.

13. (a) Find the point on the line $x + 2y = 2$ that is closest to the origin, by the method of elimination and also by Lagrange's method. Prove that your result is indeed a minimum.
(b) Repeat part (a) for the curve $x = y^2 - 2$.

14. (*Optimum soup can*) (a) Suppose that we wish to design a cylindrical soup can, out of sheet metal having a cost of α cents per unit area, with a prescribed volume V, so as to minimize the material cost. Letting x be the radius and y be the height, we have the optimization problem

$$\text{Cost}: \quad f(x,y) = \alpha(2\pi x^2 + 2\pi xy) = \min \tag{14.1}$$

subject to the constraint

$$\text{Volume}: \quad \pi x^2 y = V. \tag{14.2}$$

Solve for x and y by the method of elimination and also by the Lagrange method, and show that $x = (V/2\pi)^{1/3}$, $y = 2(V/2\pi)^{1/3}$. Verify that this solution does give a minimum.
(b) Notice from the solution to part (a) that the resulting soup can has a height that is equal to its diameter, whereas actual soup cans are a bit taller than that. Perhaps the manufacturer would not only minimize the cost, for a given volume, but would also like a height-to-diameter ratio that makes the can *look* as large as possible. Suppose the proportions that make the can look as large as possible are defined by $y = kx$, where k is an empirically determined constant. Then the optimization problem (14.1) and (14.2) can be restated as

$$\begin{aligned} a[\alpha(2\pi x^2 + 2\pi xy)] + (1-a)(y - kx)^2 &= \min, \\ \pi x^2 y &= V, \end{aligned} \tag{14.3a,b}$$

where we prescribe a such that $0 \leq a \leq 1$. For example, if we care only about minimizing the cost we set $a = 1$, if we care only about making the volume look as large as possible we set $a = 0$, if we care equally about these considerations we set $a = 0.5$, and so on. [You need not solve (14.3).] Why was it important to square the $y - kx$ term in (14.3a)? Why did we *not* need to square the square-bracketed material cost term in (14.3a)?

15. (a) Show that the application of the Lagrange method to the problem

$$f = x_1^2 + x_2^2 = \text{extremum}, \tag{15.1}$$

$$g = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 = c \tag{15.2}$$

produces the eigenvalue problem

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{15.3}$$

NOTE: Observe the correspondence between a given eigenvalue problem (15.3) and the finding of maximum and mini-

mum distances from the origin of an x_1, x_2 plane to quadric curves $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 =$ constant. As discussed in Section 11.6, $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$ is the quadratic form associated with the $\mathbf{A}$ matrix in (15.3).
(b) Sketch a representative quadric curve (15.2) in an x_1, x_2 plane and indicate on that sketch the graphical significance of the eigenvectors of $\mathbf{A}$. Explain.
(c) Find the points on the curve

$$x_1^2 + 2x_2^2 + x_1x_2 = 1$$

that are closest to the origin, and those that are farthest from the origin. Sketch the curve in the x_1, x_2 place, and show the eigenvectors of $\mathbf{A}$ and the points that you found, that are closest and farthest from the origin.
(d) Same as (c), for $x_1^2 + x_2^2 + x_1 = 1$.
(e) Same as (c), for $x_1^2 + x_2^2 - 4x_1 + 2x_2 = 2$.

16. Find the point (x, y), on the given line, that is closest to the point $(1, -2)$.

(a) $2x + y = 4$ (b) $x + y = -5$
(c) $x - 3y = 2$ (d) $x + 2y = 3$

17. Find the point (x, y, z), on the given plane, that is closest to the point $(2, 0, -1)$.

(a) $x + y + z = 4$ (b) $x - y + 2z = 4$
(c) $2x - y + z = 3$ (d) $2x + y + z = 0$

18. Find the point on the curve $y = x^{3/2}$ that is closest to the point(1,0).

19. Find the points at which $f = x + z$ is maximized and minimized, respectively, subject to the constraint $x^2 + y^2 + z^2 = 1$. Interpret the situation geometrically.

20. The constrained extremum problem

$$\begin{aligned} z &= \text{maximum}, \\ x + y + z &= 1, \\ \frac{z^2}{xy^3} &= 3 \end{aligned}$$

occurs in seeking to maximize the yield in an ammonia reactor as discussed in B. Noble, *Applications of Undergraduate Mathematics in Engineering* (New York: Macmillan, 1967), p. 75. Show that the desired maximum of z is $(17 + 4\sqrt{13})/9$.

21. (*More than one constraint*) (a) Extend the Lagrange method to cover cases that include more than one constraint. Specifically, explain how to solve the problem

$$\begin{aligned} f(x, y, z) &= \text{extremum}, \\ g_1(x, y, z) &= c_1, \\ g_2(x, y, z) &= c_2 \end{aligned}$$

by that method, and explain the logic behind each step. HINT: You should end up extremizing $f^* = f - \lambda_1 g_1 - \lambda_2 g_2$, where λ_1 and λ_2 are Lagrange multipliers.
(b) Apply that method to the problem

$$\begin{aligned} f &= x^2 + y^2 + z^2 = \text{extremum}, \\ g_1 &= x^2 + 4y^2 + 4z^2 = 4, \\ g_2 &= x + y + z = 0. \end{aligned}$$

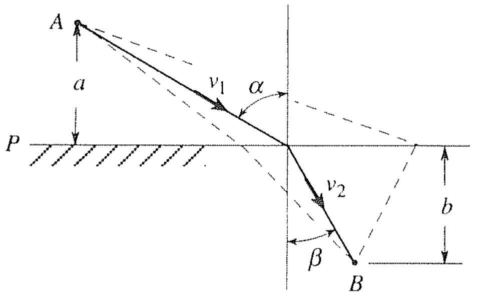
(c) Apply that method to the problem

$$\begin{aligned} f &= x + 2z = \text{extremum}, \\ g_1 &= x^2 + y^2 + z^2 = 1, \\ g_2 &= y - z = 0. \end{aligned}$$

(d) Apply that method to the problem

$$\begin{aligned} f &= x + y + z = \text{extremum}, \\ g_1 &= x^2 + y^2 + z^2 = 1, \\ g_2 &= x + y = 0. \end{aligned}$$

22. (*Fermat's principle*) (a) Let there be a light source at a given point A (shown in the first figure) in a medium in which the speed of light is v_1. Of the rays emitted from A, we are interested in the one that arrives at a given point B in a second medium in which the speed of light is v_2, the interface between the two media being the plane P. Fermat's fundamental principle of optics* states that of all possible paths



*A lawyer by profession, *Pierre de Fermat* (1601–1665) was a mathematician only by avocation. Like most mathematicians of his time, Fermat studied problems of physics as well, and his **princi-**

(such as those shown as dashed lines), the one actually followed (shown as the solid line) is the one which is such that the travel time T from A to B is a minimum. Thus, show that

$$T(\alpha,\beta) = \frac{a}{v_1 \cos\alpha} + \frac{b}{v_2 \cos\beta} = \text{minimum}, \qquad (22.1)$$

subject to the constraint that

$$g(\alpha,\beta) = a\tan\alpha + b\tan\beta = \text{constant}, \qquad (22.2)$$

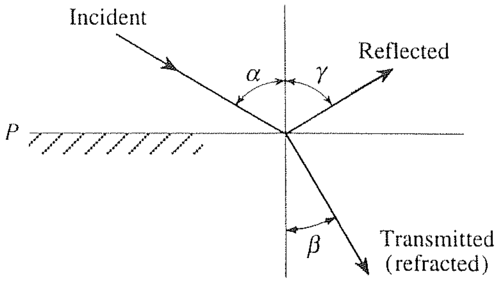
where α is the angle of incidence and β is the angle of refraction.
(b) Solving (22.1) and (22.2), derive the **law of refraction**

$$\frac{\sin\alpha}{\sin\beta} = \frac{v_1}{v_2}. \qquad (22.3)$$

(c) Besides the incident and transmitted (refracted) rays, there will also be a reflected ray, as shown in the second figure. Using the same ideas as above, derive the **law of reflection**

$$\alpha = \gamma, \qquad (22.4)$$

where γ is the angle of reflection.



NOTE: We have tacitly assumed that the paths within each medium are straight. The truth of this assumption can itself be deduced from Fermat's principle of least time.

13.8 Leibniz Rule

Recall, from the calculus, the **fundamental theorem of the integral calculus**: if f is a continuous function on the closed interval $[a, b]$, then

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t) \qquad (1)$$

for $a \leq t \leq b$. We will see, in this section and in Chapter 16, that it is also useful to be able to differentiate integrals of the form

$$I(t) = \int_{a(t)}^{b(t)} f(x,t)\,dx, \qquad (2)$$

which is more general than the integral $\int_a^t f(x)\,dx$ in (1) inasmuch as it allows for an arbitrary t dependence in both integration limits as well as in the integrand itself.

ple of least time was fundamental in the development of the science of optics. But Fermat is best remembered for his pioneering work in number theory. Of particular interest has been "Fermat's last theorem" (which, along with most of his other results, he wrote into the margin of his copy of the book *Arithmetica* by the great Alexandrian mathematician Diophantus), namely, that for $n > 2$ no integer solutions x, y, z exist for the equation $x^n + y^n = z^n$. He noted that he had "discovered a truly marvelous proof of this, which however the margin is not large enough to contain." Besides these marginal notes, and some results contained in letters to friends Fermat published but little. Thus it was, in 1665, that he perished.

To differentiate $I(t)$ it is convenient to put the sources of t dependence in a and b in evidence by expressing $I(t)$ as

$$I(t) = F(a(t), b(t), t). \tag{3}$$

That is, I is a function of the lower limit a (which, in turn, is a function of t), and of the upper limit b (which, in turn, is a function of t). The third argument in F indicates that even if a and b are fixed, I is still a function of t through the t dependence in the integrand $f(x, t)$. The form of (3) suggests chain differentiation, which step gives

$$I'(t) = \frac{\partial F}{\partial a} a'(t) + \frac{\partial F}{\partial b} b'(t) + \frac{\partial}{\partial t} \int_a^b f(x, t)\, dx, \tag{4}$$

where a, b are regarded as constants in the last term in (4), the $\partial/\partial t$ being with respect to the t dependence in the integrand. To evaluate $\partial F/\partial a$ and $\partial F/\partial b$ apply (1):

$$\frac{\partial F}{\partial a} = \frac{\partial}{\partial a} \int_a^b f(x, t)\, dx = -\frac{\partial}{\partial a} \int_b^a f(x, t)\, dx = -f(a, t) \tag{5}$$

and

$$\frac{\partial F}{\partial b} = \frac{\partial}{\partial b} \int_a^b f(x, t)\, dx = f(b, t). \tag{6}$$

Finally, it is reasonable to expect that

$$\frac{\partial}{\partial t} \int_a^b f(x, t)\, dx = \int_a^b \frac{\partial}{\partial t} f(x, t)\, dx, \tag{7}$$

that is, that we can interchange the order of the differentiation and the integration, in which case we have the result

$$\boxed{\begin{aligned} \frac{d}{dt} \int_{a(t)}^{b(t)} f(x, t)\, dx = \int_{a(t)}^{b(t)} &\frac{\partial}{\partial t} f(x, t)\, dx \\ &+ b'(t) f(b(t), t) - a'(t) f(a(t), t), \end{aligned}} \tag{8}$$

which formula is known as the **Leibniz rule.**

Under what conditions is the Leibniz rule valid?* According to the fundamental theorem (1), (5) is valid if $f(b, t)$ is a continuous function of b, that is, if f is a continuous function of its first argument. For the chain differentiation (4) to be valid, we ask that $\partial f/\partial a$ and $\partial f/\partial b$ be continuous. Finally, it can be shown that (7) holds if f and $\partial f/\partial t$ are continuous functions of x and t. Putting these results together, we state the following theorem.†

*The issue is whether the limit interchange is possible, for the integrals are limits of Riemann sums and the derivatives are limits of difference quotients. Limit interchange came up in the closure part of Section 13.3, which we urge you to review. For the validity of the interchange in (7) (which is called *differentiation under the integral sign*) to be guaranteed, we need f to be sufficiently well behaved.

†While the discussion preceding the theorem outlines the main ideas, it is not a complete proof of the theorem. For a detailed proof, see Theorem 9-38 on p.220 of T. M. Apostol's *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957).

THEOREM 13.8.1 *Leibniz Rule*
Let $a(t)$ and $b(t)$ be differentiable for $t_1 \leq t \leq t_2$, and let A and B be finite constants such that $A \leq a(t) \leq B$ and $A \leq b(t) \leq B$ on $t_1 \leq t \leq t_2$. If $f(x,t)$ and $f_t(x,t)$ are continuous on the rectangle $A \leq x \leq B$, $t_1 \leq t \leq t_2$, then the Leibniz rule (8) holds for each t on $t_1 \leq t \leq t_2$.

EXAMPLE 1. Given

$$I(t) = \int_{-t}^{t^2} \cos(tx^2)\,dx, \tag{9}$$

evaluate $I'(t)$ by means of the Leibniz rule, if that rule applies. In this case $a(t) = -t$ and $b(t) = t^2$ are both differentiable on every interval $t_1 \leq t \leq t_2$, no matter how large, and $f(x,t) = \cos(tx^2)$ and $f_t(x,t) = -x^2 \sin(tx^2)$ are continuous on every rectangle in the x,t plane, so (8) applies and gives

$$\begin{aligned} I'(t) &= -\int_{-t}^{t^2} x^2 \sin(tx^2)\,dx + (2t)\cos\left[t(t^2)^2\right] - (-1)\cos\left[t(-t)^2\right] \\ &= -\int_{-t}^{t^2} x^2 \sin(tx^2)\,dx + 2t\cos(t^5) + \cos(t^3) \end{aligned} \tag{10}$$

at each point in the x,t plane. ▮

Thus far we have not explained the *purpose* of the Leibniz rule–how it is useful to us. The two examples to follow illustrate its use, and additional applications are included among the exercises.

EXAMPLE 2. To solve the forced harmonic oscillator initial-value problem

$$mx'' + kx = f(t); \qquad x(0) = 0,\ x'(0) = 0 \tag{11}$$

one can proceed by finding homogeneous and particular solutions by the methods explained in Chapter 3 (with the particular solution found by the method of variation of parameters) or by using the Laplace transform method explained in Chapter 5. Doing so, suppose we obtain the result

$$x(t) = \frac{1}{m\omega}\int_0^t \sin\omega(t-\tau)\,f(\tau)\,d\tau, \tag{12}$$

where $\omega = \sqrt{k/m}$ is the natural frequency of the oscillator. To check that result we can use the Leibniz rule, according to which

$$\begin{aligned} x'(t) &= \frac{1}{m\omega}\int_0^t \omega\cos\omega(t-\tau)\,f(\tau)\,d\tau + \frac{1}{m\omega}(1)(\sin 0)f(t) - 0 \\ &= \frac{1}{m}\int_0^t \cos\omega(t-\tau)\,f(\tau)\,d\tau, \end{aligned} \tag{13}$$

and

$$\begin{aligned} x''(t) &= -\frac{\omega}{m}\int_0^t \sin\omega(t-\tau)\, f(\tau)\,d\tau + \frac{1}{m}(1)(\cos 0) f(t) - 0 \\ &= -\frac{\omega}{m}\int_0^t \sin\omega(t-\tau)\, f(\tau)\,d\tau + \frac{1}{m} f(t). \end{aligned} \tag{14}$$

Putting (12) and (14) into (11) does give the identity $f(t) = f(t)$. Further, it follows from (12) and (13) that the initial conditions $x(0) = 0$ and $x'(0) = 0$ are safisfied as well. ▮

EXAMPLE 3. Sometimes we can use the Leibniz rule to evaluate integrals. For example, suppose we wish to evaluate

$$I = \int_0^\infty x^2 e^{-x^2}\,dx. \tag{15}$$

If we recall the known result*

$$\int_0^\infty e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2}, \tag{16}$$

then we can evaluate I by considering instead the integral

$$J(a) = \int_0^\infty e^{-ax^2}\,dx, \tag{17}$$

for if we set $ax^2 = y^2$ (or $\sqrt{a}x = y$) in (16), then

$$J(a) = \int_0^\infty e^{-y^2}\,\frac{dy}{\sqrt{a}} = \frac{\sqrt{\pi}}{2}a^{-1/2}. \tag{18}$$

Differentiating (17) and the right-hand side of (18) with respect to a gives

$$\int_0^\infty -x^2 e^{-ax^2}\,dx = -\frac{1}{2}\frac{\sqrt{\pi}}{2}a^{-3/2} \tag{19}$$

and, finally, setting $a = 1$ in (19) gives the result

$$I = \int_0^\infty x^2 e^{-x^2}\,dx = \frac{\sqrt{\pi}}{4}. \tag{20}$$

COMMENT 1. Observe that there was no parameter in the original integral (15). To evaluate I we considered a slightly different integral J, with a suitably inserted parameter.

COMMENT 2. Actually, Theorem 13.8.1 does not suffice for the purposes of this example because of the infinite upper integration limit. That is, there does not exist the finite B that is called for in the theorem. This situation is more complicated because

$$\frac{d}{da}\int_0^\infty e^{-ax^2}\,dx = \frac{d}{da}\left(\lim_{C\to\infty}\int_0^C e^{-ax^2}\,dx\right). \tag{21}$$

*The result stated in (16) is well known and is the subject of Exercise 9 of Section 4.5.

Thus, besides the integral from 0 to C being the limit of a sequence of Riemann sums, and the derivative with respect to a being the limit of a difference quotient, we have the additional limit of C tending to ∞. The upshot is that if one or both limits in (8) are infinite ($+\infty$ or $-\infty$), then we need to impose conditions that are slightly more stringent than those in Theorem 13.8.1. Relevant theorems are to be found in books on advanced calculus, such as F. B. Hildebrand's *Advanced Calculus for Applications*, 2nd ed. (Englewood Cliffs, NJ: Prentice Hall, 1976), Sec. 7.9. We assert that (8) does hold in this example and in the exercises to follow. ▮

Closure. In summary, it is often useful to be able to differentiate a given integral with respect to a parameter without having to first carry out the integration. The formula for doing so is the Leibniz rule (8), and sufficient conditions for the validity of (8) are given in Theorem 13.8.1. Besides single integrals one is also interested in differentiating *multiple* integrals with respect to a parameter but, in the interest of brevity, we do not consider that case here.

EXERCISES 13.8

1. Apply the Leibniz rule:

(a) $\dfrac{d}{dt}\displaystyle\int_0^{t^2} \sin(tx^2)\,dx$

(b) $\dfrac{d}{dt}\displaystyle\int_3^{t} x^t \sin x\,dx$

(c) $\dfrac{d}{d\alpha}\displaystyle\int_\alpha^{2} e^{x^2}\,dx$

(d) $\dfrac{d}{d\alpha}\displaystyle\int_{-2\alpha^2}^{-\alpha} e^{\alpha x^3}\,dx$

(e) $\dfrac{d^2}{dx^2}\displaystyle\int_x^{2x} \ln(u^2+x^2)\,du$

(f) $\dfrac{d}{dy}\displaystyle\int_{y^2}^{1} x\,dx/(x^3+y^3)$

(g) $\dfrac{d^2}{da^2}\displaystyle\int_{5a}^{a^2} \cos(v^2+a^2)\,dv$

2. Derive the Taylor series of the given function $f(x)$ about $x=0$, up to and including terms of second order, using the Leibniz rule to obtain $f'(x)$ and $f''(x)$.

(a) $\int_0^x e^{-t^2}\,dt$ (b) $\int_{-x}^{\cos x} dt/(t^3+1)$

(c) $\int_0^{2\sin x} \ln(t^2+1)\,dt$ (d) $\int_0^{1+2x} e^{-xt^2}\,dt$

(e) $\int_x^{2x} \sin(xt^2)\,dt$ (f) $\int_{-x}^{3\sin x} \cos(xt^2)\,dt$

3. Show, by repeated differentiation of the formula

$$\int_0^\infty e^{-ax}\,dx = \frac{1}{a},$$

that $\int_0^\infty x^n e^{-x}\,dx = n!$ for $n = 0, 1, 2, \ldots$.

4. Evaluate $\int_0^\infty x^4 e^{-x^2}\,dx$, using any of the formulas found in Example 3.

5. (a) To evaluate

$$I = \int_0^\infty \frac{(\ln x)^2}{1+x^3}\,dx,$$

differentiate the formula

$$\int_0^\infty \frac{x^a}{1+x^3}\,dx = \frac{\pi}{3\sin\left(\frac{a+1}{3}\pi\right)} \tag{5.1}$$

a suitable number of times. Thus, show that $I = 10\pi^3/(81\sqrt{3})$. (b) Use (5.1) to evaluate the integral

$$I = \int_0^\infty \frac{x\ln x}{1+x^3}\,dx.$$

HINT: Recall that $d(x^a)/da = x^a \ln x$.

6. Show that

(a) $y(x) = \frac{1}{6}\int_a^x (x-t)^3 f(t)\,dt$ satisfies the initial-value problem $y''''(x) = f(x);\ \ y(a) = y'(a) = y''(a) =$

$y'''(a) = 0$
(b) $y(x) = (1/x)\int_a^x (x-t)f(t)\,dt$ satisfies the initial-value problem $(xy)'' = f(x)$; $y(a) = y'(a) = 0$
(c) $y(x) = e^x + \int_0^x t^2 \cosh(x-t)\,dt$ satisfies the intial-value problem $y'' - y = 2x$; $y(0) = y'(0) = 1$

7. Evaluate $\int_0^1 x^{0.7} \ln x\,dx$ by differentiating the known formula

$$\int_0^1 x^a\,dx = \frac{1}{a+1}. \qquad (a > -1)$$

8. Show that

$$\int_0^1 \frac{x^3-1}{\ln x}\,dx = \ln 4.$$

HINT: Considering

$$I(a) = \int_0^1 \frac{x^a - 1}{\ln x}\,dx, \qquad (a \geq 0)$$

show that $I'(a) = 1/(a+1)$ with the "initial condition" $I(0) = 0$. Solve for $I(a)$ and set $a = 3$.

9. The conditions in Theorem 13.8.1 are sufficient, but not necessary. To illustrate this point, show that Leibniz differentiation of the integral

$$I(t) = \int_0^t \frac{e^{\sqrt{x}}}{\sqrt{x}}\,dx \qquad (t > 0)$$

does give the correct result even though the integrand $f(x,t) = e^{\sqrt{x}}/\sqrt{x}$ is not continuous at $x = 0$.

10. Consider a one-dimensional fluid flow, say flow in a channel of cross-sectional area A, with velocity $v(x,t)$ m/sec and mass density $\sigma(x,t)$ grams/m^3; x is measured positive downstream and t is the time. Consider a "control volume" $a(t) \leq x \leq b(t)$ that drifts with the fluid; i.e., $da/dt = v(a(t),t)$ and $db/dt = v(b(t),t)$. Then show that the principle of conservation of mass (i.e., that the amount of mass within the control volume remains constant) can be expressed as

$$\begin{aligned}\frac{d}{dt}&\int_{a(t)}^{b(t)} \sigma(x,t)A\,dx \\ &= A\int_{a(t)}^{b(t)} \left[\frac{\partial\sigma}{\partial t} + \frac{\partial}{\partial x}(v\sigma)\right] dx = 0.\end{aligned}$$

11. Show that

$$\text{(a)} \quad u(x,t) = \int_0^1 f(\xi)\frac{e^{-(\xi-x)^2/(4\alpha^2 t)}}{2\alpha\sqrt{\pi t}}\,d\xi \qquad (11.1)$$

$(-\infty < x < \infty, t > 0)$ satisfies the PDE (partial differential equation) $\alpha^2 u_{xx} = u_t$, known as the (one- dimensional) **diffusion equation**, where α^2 is a constant known as the diffusivity of the medium, and where f is a prescribed continuous function

$$\text{(b)}\ y(x,t) = \frac{F(x-ct) + F(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} G(\alpha)\,d\alpha \qquad (11.2)$$

$(-\infty < x < \infty, t > 0)$ satisfies the PDE $c^2 y_{xx} = y_{tt}$, known as the (one-dimensional) **wave equation**, as well as the initial conditions $y(x,0) = F(x)$ and $y_t(x,0) = G(x)$, where $F(x)$ is differentiable and $G(x)$ is continuous

$$\text{(c)} \quad u(x,y) = \frac{y}{\pi}\int_{-1}^1 \frac{f(\xi)}{(\xi-x)^2 + y^2}\,d\xi \qquad (11.3)$$

$(-\infty < x < \infty,\ y > 0)$ satisfies the PDE $u_{xx} + u_{yy} = 0$, known as the (two-dimensional) **Laplace equation**, where f is a prescribed continuous function.

12. (*Double and triple integrals*) The Leibniz rule (8) can be extended to multiple integrals. For our purposes it will suffice to limit that extension to the case of double and triple integrals over regions that are independent of the parameter. In the one-dimensional version (8), if $a(t)$ and $b(t)$ are constants then we have

$$\frac{d}{dt}\int_a^b f(x,t)\,dx = \int_a^b \frac{\partial}{\partial t}f(x,t)\,dx, \qquad (12.1)$$

i.e., the derivative of the integral equals the integral of the derivative. In two and three dimensions the analogs of (12.1) are

$$\frac{d}{dt}\iint_{\mathcal{R}} f(x,y,t)\,dA = \iint_{\mathcal{R}} \frac{\partial}{\partial t}f(x,y,t)\,dA \qquad (12.2)$$

and

$$\frac{d}{dt}\iiint_{\mathcal{R}} f(x,y,z,t)\,dV = \iiint_{\mathcal{R}} \frac{\partial}{\partial t}f(x,y,z,t)\,dV. \qquad (12.3)$$

Sufficient conditions for the validity of (12.2) and (12.3) are that f and $\partial f/\partial t$ are continuous on the closed region $\mathcal{R}$ over the t interval of interest.

(a) Use (12.2) to show that

$$u(x,y) = \int_0^1\int_0^1 \ln\left[(x-\xi)^2 + (y-\eta)^2\right] f(\xi,\eta)\,d\xi\,d\eta \qquad (12.4)$$

satisfies the PDE $u_{xx} + u_{yy} = 0$, known as the (two-dimensional) **Laplace equation**, at all points (x,y) outside

the square $0 \le x \le 1,\ 0 \le y \le 1$, where f is any prescribed function that is continuous on that square.
(b) Use (12.3) to show that

$$u(x, y, z) = \int_0^1 \int_0^1 \int_0^1 \frac{f(\xi, \eta, \zeta)\, d\xi\, d\eta\, d\zeta}{\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}}$$

satisfies the (three-dimensional) **Laplace equation** $u_{xx} + u_{yy} + u_{zz} = 0$ at all points (x, y, z) outside the cube $0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 1$, where f is any prescribed function that is continuous on that cube.

Chapter 13 Review

For the most part, Sections 13.1–13.5.1 are devoted to a review of the necessary preliminaries from the calculus, including point set theory, limits and continuity, composite functions and chain differentiation, Taylor's formula, and Taylor series.

In Section 13.5.2 we extend Taylor's formula and Taylor series to functions of more than one variable, such as $f(x, y)$, by parametrizing a line from the initial point (a, b) to the final point (x_0, y_0) by

$$x = a + (x_0 - a)t, \qquad y = b + (y_0 - b)t, \qquad (0 \le t \le 1)$$

because then $f(x, y) = f(x(t), y(t))$ reduces to a function $F(t)$ of the single variable t. The single-variable Taylor expansion of $F(t)$ about $t = 0$ leads us to the desired Taylor expansion of $f(x, y)$ about (a, b). An alternative and readily implemented approach is given in Exercise 11 of that section, namely, simply expanding in one variable at a time.

In Section 13.6 we consider relations such as $f(x, y) = 0$ on x and y, and the relations $f(x, y, u, v) = 0$ and $g(x, y, u, v) = 0$ on x, y, u, and v, and inquire as to whether such relations imply the existence of implicit functions $y(x)$, and $u(x, y)$ and $v(x, y)$, respectively. The chief result is the Implicit Function Theorem 13.6.2, which emphasizes the importance of the nonvanishing of the Jacobian determinant $\dfrac{\partial(f_1, \dots, f_n)}{\partial(u_1, \dots, u_n)}$ at the point in question, if the relations

$$\begin{gathered} f_1(x_1, \dots, x_n, u_1, \dots, u_n) = 0, \\ \vdots \\ f_n(x_1, \dots, x_n, u_1, \dots, u_n) = 0 \end{gathered}$$

are to imply the existence of inverse functions $u_1(x_1, \dots, x_n), \dots, u_n(x_1, \dots, x_n)$. We also find that the result

$$\frac{dy}{dx}\frac{dx}{dy} = 1,$$

for functions $y(x)$ and $x(y)$, generalizes to

$$\frac{\partial(x, y)}{\partial(u, v)}\frac{\partial(u, v)}{\partial(x, y)} = 1, \qquad \frac{\partial(x, y, z)}{\partial(u, v, w)}\frac{\partial(u, v, w)}{\partial(x, y, z)} = 1,$$

and so on, and that the chain rule

$$\frac{du}{dx}\frac{dx}{ds} = \frac{du}{ds}$$

for the function $u(x(s))$ generalizes to

$$\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(r,s)} = \frac{\partial(u,v)}{\partial(r,s)},$$
$$\frac{\partial(u,v,w)}{\partial(x,y,z)}\frac{\partial(x,y,z)}{\partial(r,s,t)} = \frac{\partial(u,v,w)}{\partial(r,s,t)},$$

and so on.

In Section 13.7 we review the theory of maxima and minima for functions of a single variable before turning to the unconstrained and constrained extremization of functions of more than one variable, considering only local extrema at interior points of the domain of the function being extremized. If $f(x_1, \dots, x_n)$ is C^1 in some neighborhood of the point of extremum, $\mathbf{X}$, then the extremum condition $df = 0$ implies that we must have

$$\frac{\partial f}{\partial x_1} = \cdots = \frac{\partial f}{\partial x_n} = 0$$

at $\mathbf{X}$, which is called a critical point of f. To determine if f has a maximum, minimum, or saddle at $\mathbf{X}$, we use Taylor's formula about $\mathbf{X}$, through second-order terms because the first-order terms vanish by virtue of (7). The second-order terms constitute a quadratic form and everything hinges on the eigenvalues of the matrix $\mathbf{A} = \{f_{x_i x_j}(\mathbf{X})\}$ of that quadratic form: if all of $\mathbf{A}$'s eigenvalues are negative then f has a local maximum at $\mathbf{X}$, if all are positive then f has a local minimum there, and if they are mixed in sign then f has a saddle there.

For the extremization of $f(x_1, \dots, x_n)$ in the presence of one or more constraints of the form $g(x_1, \dots, x_n) =$ constant, the extremum condition $df = 0$ still holds, but no longer implies (7). We present two methods of solving such problems, elimination and the method of Lagrange multipliers.

Finally, in Section 13.8 we derive the Leibniz rule

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)\,dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x,t)\,dx + b'(t)f(b(t),t) - a'(t)f(a(t),t)$$

for the differentiation of an integral that depends on a parameter t, which rule amounts to a generalization of the familiar fundamental theorem of the integral calculus,

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t).$$

Chapter 14

Vectors in 3-Space

14.1 Introduction

Vectors have already been studied in some detail in Chapter 9. However, except for Sections 9.1–9.3 that discussion focus mostly on n-tuple vectors in n-space. Here, however, we return to "arrow vectors" in physical 3-space and introduce a number of additional concepts, including the cross product of two vectors, combinations of dot and cross products, the differentiation of vector functions, and polar, cylindrical, and spherical coordinate systems.

Though each is of importance in its own right, Chapters 13-15 can be thought of as preparing the way for the especially important Chapter 16 on Scalar and Vector Field Theory.

14.2 Dot and Cross Product

It will be assumed here that you have studied Sections 9.1–9.3. By way of a brief review, recall that in Section 9.2 we introduced the notion of a vector in terms of a directed magnitude; we defined the magnitude or *norm* $\|\mathbf{u}\|$ of a vector $\mathbf{u}$, the *addition* of vectors $\mathbf{u}+\mathbf{v}$, the *scalar multiplication* $\alpha\mathbf{u}$, the *negative inverse* $-\mathbf{u}$, and the zero vector $\mathbf{0}$. And in Section 9.3 we defined the *dot product* $\mathbf{u}\cdot\mathbf{v}$ between any two vectors $\mathbf{u}$ and $\mathbf{v}$ as

$$\boxed{\mathbf{u}\cdot\mathbf{v}=\begin{cases}\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta & \text{if } \mathbf{u},\ \mathbf{v}\neq\mathbf{0},\\ 0 & \text{if } \mathbf{u}=\mathbf{0} \text{ or } \mathbf{v}=\mathbf{0},\end{cases}} \tag{1}$$

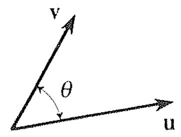


Figure 1. The angle θ in (1).

where θ is the interior angle ($0\le\theta\le\pi$) between $\mathbf{u}$ and $\mathbf{v}$ (Fig. 1). If $\mathbf{u}=\mathbf{v}\neq\mathbf{0}$, then $\theta=0$, and (1) gives $\mathbf{u}\cdot\mathbf{u}=\|\mathbf{u}\|^2$, which result holds even if $\mathbf{u}=\mathbf{v}=\mathbf{0}$. Solving for $\|\mathbf{u}\|$, we obtain the relationship

$$\|\mathbf{u}\|=\sqrt{\mathbf{u}\cdot\mathbf{u}}, \tag{2}$$

noted in Section 9.3, between the norm and the dot product. The dot product, defined by (1), admits the following properties:

$$\begin{aligned} \mathbf{u}\cdot\mathbf{v} &= \mathbf{v}\cdot\mathbf{u} && \text{(commutativity)} && \text{(3a)}\\ (\alpha\mathbf{u})\cdot\mathbf{v} &= \alpha(\mathbf{u}\cdot\mathbf{v}) && \text{(associtivity)} && \text{(3b)}\\ (\mathbf{u}+\mathbf{v})\cdot\mathbf{w} &= \mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w} && \text{(distributivity)} && \text{(3c)} \end{aligned}$$

for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and scalar α.* Of these properties, (3b) and (3c) are equivalent to the single statement that

$$(\alpha\mathbf{u}+\beta\mathbf{v})\cdot\mathbf{w} = \alpha(\mathbf{u}\cdot\mathbf{w})+\beta(\mathbf{v}\cdot\mathbf{w}) \qquad \text{(linearity)} \qquad (4)$$

for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and any scalars α, β, and the property (4) is known as *linearity*.

These properties permit us to manipulate dot products as if by ordinary scalar algebra. For example, just as

$$(2a+b)(4a-3b) = 8a^2 - 2ab - 3b^2,$$

where a, b are scalars, it is equally true that

$$\begin{aligned} (2\mathbf{u}+\mathbf{v})\cdot(4\mathbf{u}-3\mathbf{v}) &= 8\mathbf{u}\cdot\mathbf{u} - 2\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v}\\ &= 8\left\|\mathbf{u}\right\|^2 - 2\mathbf{u}\cdot\mathbf{v} - 3\left\|\mathbf{v}\right\|^2 \end{aligned}$$

since

$$\begin{aligned} (2\mathbf{u}+\mathbf{v})\cdot(4\mathbf{u}-3\mathbf{v}) &= 2\mathbf{u}\cdot(4\mathbf{u}-3\mathbf{v}) + \mathbf{v}\cdot(4\mathbf{u}-3\mathbf{v}) && \text{by (4)}\\ &= 2(4\mathbf{u}-3\mathbf{v})\cdot\mathbf{u} + (4\mathbf{u}-3\mathbf{v})\cdot\mathbf{v} && \text{by (3a)}\\ &= 2(4\mathbf{u}\cdot\mathbf{u} - 3\mathbf{v}\cdot\mathbf{u} + 4\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v} && \text{by (4)}\\ &= 8\mathbf{u}\cdot\mathbf{u} - 6\mathbf{u}\cdot\mathbf{v} + 4\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v} && \text{by (3a)}\\ &= 8\left\|\mathbf{u}\right\|^2 - 2\mathbf{u}\cdot\mathbf{v} - 3\left\|\mathbf{v}\right\|^2. \end{aligned}$$

Before leaving the dot product, observe from (1) that if $\mathbf{u}$ and $\mathbf{v}$ are nonzero, and $\theta = \pi/2$, then $\mathbf{u}$ and $\mathbf{v}$ are **perpendicular** to each other, and $\mathbf{u}\cdot\mathbf{v} = 0$. The converse, however, is not true for if $\mathbf{u}\cdot\mathbf{v} = 0$, then either $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, *or* at least one of $\mathbf{u}$ and $\mathbf{v}$ is zero. Thus, we introduce a separate term, *orthogonal*: $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if $\mathbf{u}\cdot\mathbf{v} = 0$. Perpendicularity implies orthogonality, but only if $\mathbf{u}$ and $\mathbf{v}$ are both nonzero does orthogonality imply perpendicularity.

Further, we say that a *set* of vectors $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is an **orthogonal set** if every vector in the set is orthogonal to every other vector in the set:

$$\mathbf{v}_i\cdot\mathbf{v}_j = 0 \quad \text{if } i \neq j. \qquad (5)$$

If, in addition, each $\mathbf{v}_j$ is scaled or normalized so as to have unit length, then

$$\mathbf{v}_i\cdot\mathbf{v}_j = \begin{cases} 1, & i = j\\ 0, & i \neq j \end{cases} \qquad (6)$$

*In Part III (Chapters 13–16), all scalars are *real* numbers.

since $\mathbf{v}_j \cdot \mathbf{v}_j = \|\mathbf{v}_j\|^2 = 1$, and we say that $\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ is an **orthonormal set**, namely, both orthogonal and normalized. Such sets are used extensively in subsequent sections.

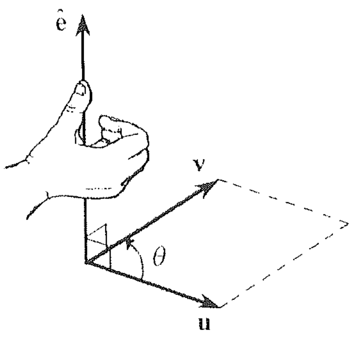


Figure 2. Cross product.

Besides the dot product, we now introduce another kind of product between two vectors, the **cross product** of $\mathbf{u}$ and $\mathbf{v}$, denoted as $\mathbf{u} \times \mathbf{v}$, and defined as[†]

$$\boxed{\mathbf{u} \times \mathbf{v} = \begin{cases} \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}} & \text{if } \mathbf{u}, \mathbf{v} \neq \mathbf{0},\ \ \theta \neq 0, \text{ and } \theta \neq \pi, \\ \mathbf{0} & \text{if } \mathbf{u} = \mathbf{0},\ \ \mathbf{v} = \mathbf{0},\ \ \theta = 0, \text{ or } \theta = \pi, \end{cases}} \qquad (7)$$

where θ is again the interior angle between $\mathbf{u}$ and $\mathbf{v}$, and $\hat{\mathbf{e}}$ is a unit vector normal to the plane of $\mathbf{u}$ and $\mathbf{v}$, directed such that the vectors $\mathbf{u}$, $\mathbf{v}$, $\hat{\mathbf{e}}$ form a *right-handed system*. That is, curling the other four fingers of our right hand from $\mathbf{u}$ (the first vector in $\mathbf{u} \times \mathbf{v}$) into $\mathbf{v}$ (the second vector in $\mathbf{u} \times \mathbf{v}$), our thumb points in the $\hat{\mathbf{e}}$ direction (Fig. 2).

Observe, first, that while the dot product $\mathbf{u} \cdot \mathbf{v}$ is a scalar, the cross product $\mathbf{u} \times \mathbf{v}$ is a vector. For that reason, $\mathbf{u} \cdot \mathbf{v}$ is also called the **scalar product** of $\mathbf{u}$ and $\mathbf{v}$, and $\mathbf{u} \times \mathbf{v}$ is also called the **vector product** of $\mathbf{u}$ and $\mathbf{v}$. (Some authors use the notation "$\mathbf{uv}$" in place of our $\mathbf{u} \cdot \mathbf{v}$, and "$\mathbf{u} \wedge \mathbf{v}$" in place of our $\mathbf{u} \times \mathbf{v}$.)

EXAMPLE 1. Let $\mathbf{u}$, $\mathbf{v}$ be as shown in Fig. 3: $\|\mathbf{u}\| = 5$, $\|\mathbf{v}\| = 4$, and $\theta = 60°$. Then $\mathbf{u} \times \mathbf{v} = (5)(4)(\sin 60°)\hat{\mathbf{e}} = (10\sqrt{3})\hat{\mathbf{e}}$ as sketched in the figure; $\mathbf{v} \times \mathbf{u}$ is of the same magnitude as $\mathbf{u} \times \mathbf{v}$ but points in the opposite direction. ▮

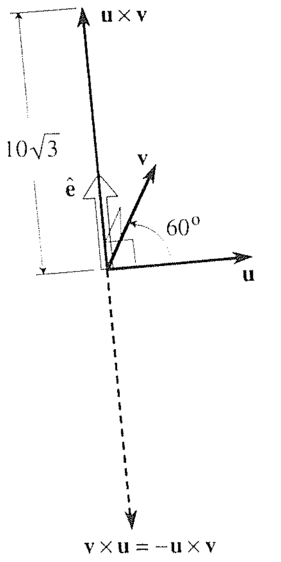


Figure 3. $\mathbf{u} \times \mathbf{v}$ and $\mathbf{v} \times \mathbf{u}$.

In contrast with the commutativity property of the dot product, namely $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$, we have $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$ for the cross product since

$$\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}}_1, \qquad \mathbf{v} \times \mathbf{u} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}}_2,$$

where, according to the right-hand rule, $\hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_1$. The property $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$ is known as *anticommutativity* and is illustrated in Fig. 3. While commutativity fails to hold, associative and distributive properties do hold:

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= -(\mathbf{v} \times \mathbf{u}) && (\textit{anti}\text{commutativity}) && (8a) \\ (\alpha\mathbf{u}) \times \mathbf{v} &= \alpha(\mathbf{u} \times \mathbf{v}) && (\text{associativity}) && (8b) \\ (\mathbf{u} + \mathbf{v}) \times \mathbf{w} &= (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w}) && (\text{distributivity}) && (8c) \end{aligned}$$

Of these properties, (8b) and (8c) are equivalent to the single statement

$$(\alpha\mathbf{u} + \beta\mathbf{v}) \times \mathbf{w} = \alpha(\mathbf{u} \times \mathbf{w}) + \beta(\mathbf{v} \times \mathbf{w}) \qquad (\text{linearity}) \qquad (9)$$

[†] You may wonder why we need to spell out, in (7), that $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if $\theta = 0$ or $\theta = \pi$ since, after all, (7) *already* seems to say that $\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}} = 0\,\hat{\mathbf{e}} = \mathbf{0}$ (since $\sin 0 = 0$ and $\sin\pi = 0$). The catch is that $\hat{\mathbf{e}}$ is *undefined* if $\theta = 0$ or $\theta = \pi$ since then the $\mathbf{u}$, $\mathbf{v}$ vectors fail to define a unique plane.

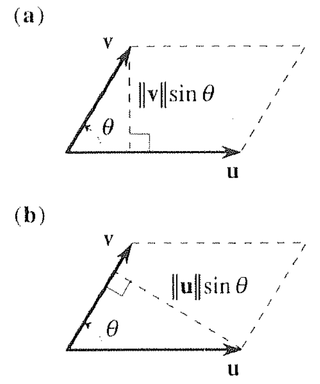


Figure 4. Area significance of $\|\mathbf{u} \times \mathbf{v}\|$.

for any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ and scalars α, β, and the property (9) is known as *linearity*. Proof of (8b) and (8c) is left for the exercises.

Besides these properties there is a geometrical significance to be noted. Specifically, *the magnitude of* $\mathbf{u} \times \mathbf{v}$ *is equal to the area of the* $\mathbf{u}, \mathbf{v}$ *parallelogram* as shown in Fig. 4 (in which $\mathbf{u}$ and $\mathbf{v}$ lie in the plane of the paper). For the area (shaded) is the altitude $\|\mathbf{v}\| \sin\theta$ times the base $\|\mathbf{u}\|$ as labeled in Fig. 4a, or, equivalently, the altitude $\|\mathbf{u}\| \sin\theta$ times the base $\|\mathbf{v}\|$ as labeled in Fig. 4b. Let us state this result as an equation, for reference:

$$\boxed{\|\mathbf{u} \times \mathbf{v}\| = \text{area of } \mathbf{u}, \mathbf{v} \text{ parallelogram.}} \tag{10}$$

Recall from Example 1 in Section 9.3 that a simple physical example of the dot product involves the work W done when a body undergoes a linear displacement from an initial point A to a final point B under the action of a constant force $\mathbf{F}$. Specifically, if we denote the displacement vector as $\mathbf{AB}$, then $W = \mathbf{F} \cdot \mathbf{AB}$. Likewise there is, again from the subject of mechanics, a simple physical example of the cross product, namely, the moment of a force.

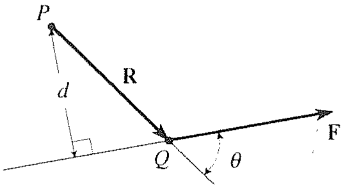


Figure 5. Moment of $\mathbf{F}$ about P.

EXAMPLE 2. *Moment of a Force.* In mechanics, one defines the *moment* M of a force $\mathbf{F}$, about a point P, as the product $\|\mathbf{F}\|\, d$, where d is the perpendicular distance from P to the line of action L of $\mathbf{F}$, as shown in Fig. 5. Besides the scalar quantity M, we can make the moment into a vector by using the right-hand rule. Specifically, one defines the vector moment $\mathbf{M}$ of $\mathbf{F}$, about P, as

$$\mathbf{M} \equiv \mathbf{R} \times \mathbf{F}, \tag{11}$$

where $\mathbf{R} = \mathbf{PQ}$ is a vector from P to the tail of $\mathbf{F}$ (Fig. 5) for then

$$\mathbf{M} = \|\mathbf{R}\|\,\|\mathbf{F}\| \sin\theta\, \hat{\mathbf{e}} = \|\mathbf{F}\|\,(\|\mathbf{R}\| \sin\theta)\, \hat{\mathbf{e}} = \|\mathbf{F}\|\, d\, \hat{\mathbf{e}} = M\hat{\mathbf{e}},$$

where $\hat{\mathbf{e}}$ is a unit vector perpendicular to the plane of $\mathbf{R}$ and $\mathbf{F}$ and in the direction dictated by the right-hand rule. For instance, if $\mathbf{R}$ and $\mathbf{F}$ in Fig. 5 are in the plane of the paper, then $\hat{\mathbf{e}}$ is directed perpendicular to the paper and toward the reader.

Notice that $\mathbf{M}$ is independent of the position of the point Q on the line of action L. For suppose that Q_1 and Q_2 are any two points on L (Fig. 6). Then

$$\begin{aligned} \mathbf{PQ}_1 \times \mathbf{F} - \mathbf{PQ}_2 \times \mathbf{F} &= (\mathbf{PQ}_1 - \mathbf{PQ}_2) \times \mathbf{F} && \text{by (8c),} \\ &= \mathbf{Q}_2\mathbf{Q}_1 \times \mathbf{F} && \text{by Fig. 6} \\ &= 0 \end{aligned}$$

since the angle between $\mathbf{Q}_2\mathbf{Q}_1$ and $\mathbf{F}$ is either 0 or π. Thus, $\mathbf{PQ}_1 \times \mathbf{F} = \mathbf{PQ}_2 \times \mathbf{F}$, as claimed. ∎

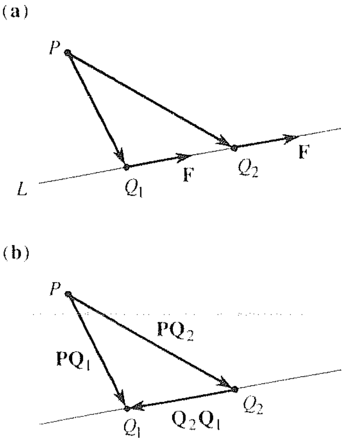


Figure 6. Location of Q.

Closure. Whereas a dot product can be defined for *any* vector space, the cross product is specific to 3-space. In this section we recalled the dot product (1), from Section 9.3, and introduce the cross product (7). Both operations are linear, but whereas the dot product is commutative ($\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$), the cross product is anti-commutative ($\mathbf{u}\times\mathbf{v} = -\mathbf{v}\times\mathbf{u}$).

EXERCISES 14.2

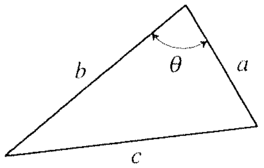
1. If $\mathbf{u}\times\mathbf{v} = 2\mathbf{u}$, what can one conclude about $\mathbf{u}$? About $\mathbf{v}$? Explain.

2. If $\mathbf{u}\times\mathbf{v} = \mathbf{0}$ and $\mathbf{u}\cdot\mathbf{v} = 0$, must $\mathbf{u} = \mathbf{0}$ and/or $\mathbf{v} = \mathbf{0}$? Explain.

3. Use the dot product to derive the *law of cosines*,

$$c^2 = a^2 + b^2 - 2ab\cos\theta,$$

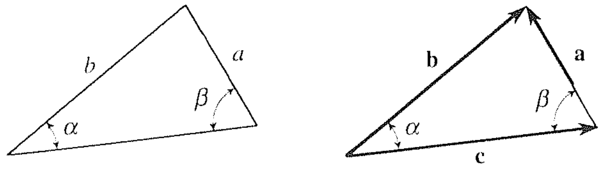
where c, a, b, θ are shown in the figure. HINT: Regard the sides of the triangle as vectors, oriented such that $\mathbf{c} = \mathbf{a} - \mathbf{b}$, and note that $\mathbf{c}\cdot\mathbf{c} = (\mathbf{a}-\mathbf{b})\cdot(\mathbf{a}-\mathbf{b})$.



4. Use the cross product to derive the *law of sines*,

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta},$$

where a, b, α, β are as shown in part (a) of the accompanying figure. HINT: Regard the sides of the triangle as vectors, as in part (b), and cross a suitable vector into both sides of the equation $\mathbf{b} - \mathbf{c} = \mathbf{a}$.



(a) (b)

5. Derive the **Lagrange identity**

$$\|\mathbf{u}\times\mathbf{v}\|^2 = \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 - (\mathbf{u}\cdot\mathbf{v})^2. \tag{5.1}$$

6. Why did we bother to include the parentheses in the right-hand side of (8c) but not in the analogous equation (3c)?

7. If $\mathbf{u}\cdot\mathbf{v} = 0$ and $\mathbf{u}\times\mathbf{v} \equiv \mathbf{w}$ is $\{\mathbf{u},\mathbf{v},\mathbf{w}\}$ necessarily an orthogonal set? Explain.

8. Is $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$ necessarily equal to $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$? Prove or disprove.

9. Prove the associative law (8b). [Proof of the distributive law (8c) is left for the exercises in Section 14.3.]

10. (a) Prove that

$$(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n)\cdot\mathbf{v} =$$
$$\alpha_1(\mathbf{u}_1\cdot\mathbf{v}) + \alpha_2(\mathbf{u}_2\cdot\mathbf{v}) + \cdots + \alpha_n(\mathbf{u}_n\cdot\mathbf{v}),$$

namely, that (4) holds even if $n > 2$.
(b) Prove that

$$(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n)\times\mathbf{v} =$$
$$\alpha_1(\mathbf{u}_1\times\mathbf{v}) + \alpha_2(\mathbf{u}_2\times\mathbf{v}) + \cdots + \alpha_n(\mathbf{u}_n\times\mathbf{v}),$$

namely, that (9) holds even if $n > 2$.

11. (*Linear dependence*) Prove that vectors $\mathbf{u}, \mathbf{v}$ are linearly dependent if and only if $\mathbf{u}\times\mathbf{v} = \mathbf{0}$.

12. Prove that if $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors, and $\mathbf{u}\times\mathbf{v} = \mathbf{0}$, then $\mathbf{u} = \alpha\mathbf{v}$ for some nonzero constant α.

14.3 Cartesian Coordinates

In Sections 9.2 and 14.2 we introduce vectors in 3-space and define the norm of a vector, the multiplication of a vector by a scalar, addition and subtraction of vectors, and the dot and cross products of two vectors. These definitions are *intrinsic* inasmuch as they contain no reference whatsoever to any coordinate system.

In applications, however, our computations are often carried out more conveniently if we introduce a reference coordinate system and a corresponding set of base vectors. Since the operations listed above are defined intrinsically, it follows that the answers obtained will be the same whether we choose one coordinate system and set of base vectors or another. Thus, the choice is a matter of preference and convenience.

Although there are many coordinate systems from which to choose, we limit our attention here to those that we regard as the most important: *Cartesian*, *cylindrical*, and *spherical*. In the present section we consider the Cartesian case.

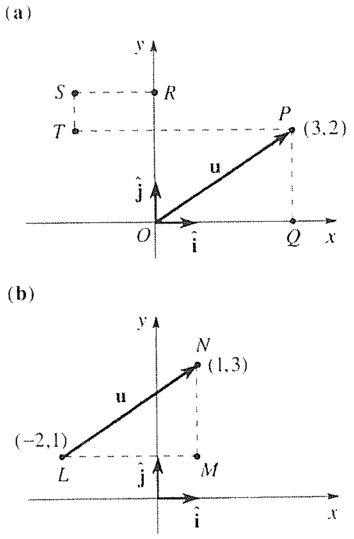


Figure 1. Cartesian x, y system.

Beginning with 2-space, for simplicity, we adopt the rectangular Cartesian x, y coordinate system shown in Fig. 1, together with the reference vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}$, where $\hat{\mathbf{i}}$ is a unit vector in the positive x direction, and $\hat{\mathbf{j}}$ is a unit vector in the positive y direction. We have placed $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ with their tails at the origin, but one can place them wherever one wishes, as long as their magnitudes and directions are preserved. Observe that $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}\right\}$ constitutes an orthonormal set since

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{i}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{j}} = (1)(1)\cos 0^\circ = 1 \qquad \text{and} \qquad \hat{\mathbf{i}} \cdot \hat{\mathbf{j}} = (1)(1)\cos 90^\circ = 0. \qquad (1)$$

Now, consider the vector $\mathbf{u}$ in Fig. 1a; whether $\mathbf{u}$ is being used as a *position vector* to locate the point P, namely the point $(3, 2)$ in the x, y plane or to designate a certain force vector * or a velocity vector or whatever, is not important here. We can express $\mathbf{u}$ as a linear combination of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ as follows:

$$\mathbf{u} = \mathbf{OQ} + \mathbf{QP} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}.$$

The end result, $\mathbf{u} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, is independent of the fact that the $\mathbf{u}$ vector happens to spring from the origin; if we move its tail to $(-2, 1)$, say, as in Fig. 1b, then

$$\mathbf{u} = \mathbf{LM} + \mathbf{MN} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}},$$

which result is the same as above.

More generally, let $\mathbf{u}$ be a vector from the origin to *any* point (a, b) [or, equivalently, from (x_0, y_0) to $(x_0 + a, y_0 + b)$]. By the same reasoning as above, we can express $\mathbf{u}$ in terms of $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ as

$$\mathbf{u} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}. \qquad (2)$$

Is (2) unique? For example, breaking OP (in Fig. 1a) into OQP gave $\mathbf{u} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$; but if we had broken it into $ORSTP$, instead, might we have obtained a different result, such as $\mathbf{u} = 4\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$? No. To prove that the expression (2) of a given vector$\mathbf{u}$, as a linear combination of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$, is indeed unique, suppose that $\mathbf{u}$ also admits an expansion

$$\mathbf{u} = a'\hat{\mathbf{i}} + b'\hat{\mathbf{j}}. \qquad (3)$$

*For example, if we choose a scale of 1 newton per unit length, then $\mathbf{u}$, in Fig. 1a, would represent a force of $\sqrt{13}$ newtons, in the direction OP.

Subtracting equals from equals, namely, (3) from (2) yields

$$(a-a')\hat{\mathbf{i}}+(b-b')\hat{\mathbf{j}}=\mathbf{0}. \tag{4}$$

Dotting both sides of (4) with $\hat{\mathbf{i}}$, then with $\hat{\mathbf{j}}$, and using (1), we find that $a-a'=0$ and $b-b'=0$ so that $a=a'$ and $b=b'$. Consequently, (3) is necessarily identical to (2) so (2) is unique, as claimed.

Similarly for 3-space: Suppose that we adopt a Cartesian x, y, z coordinate system, with reference vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ (Fig. 2), where

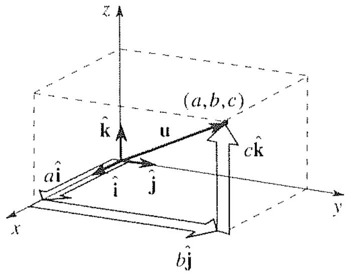


Figure 2. Cartesian coordinates and $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$.

$$\hat{\mathbf{i}}\cdot\hat{\mathbf{i}}=\hat{\mathbf{j}}\cdot\hat{\mathbf{j}}=\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}=1, \tag{5a}$$

$$\hat{\mathbf{i}}\cdot\hat{\mathbf{j}}=\hat{\mathbf{i}}\cdot\hat{\mathbf{k}}=\hat{\mathbf{j}}\cdot\hat{\mathbf{k}}=0. \tag{5b}$$

It follows from (5) that $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right\}$ is an orthonormal set. Furthermore,

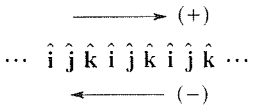
$$\hat{\mathbf{i}}\times\hat{\mathbf{j}}=\left\|\hat{\mathbf{i}}\right\|\left\|\hat{\mathbf{j}}\right\|\sin 90^\circ\,\hat{\mathbf{k}}=\hat{\mathbf{k}}. \tag{6}$$

Working out all of the various cross products in this manner, we find that

$$\hat{\mathbf{i}}\times\hat{\mathbf{i}}=\hat{\mathbf{j}}\times\hat{\mathbf{j}}=\hat{\mathbf{k}}\times\hat{\mathbf{k}}=\mathbf{0}, \tag{7a}$$

$$\hat{\mathbf{i}}\times\hat{\mathbf{j}}=\hat{\mathbf{k}},\qquad \hat{\mathbf{j}}\times\hat{\mathbf{k}}=\hat{\mathbf{i}},\qquad \hat{\mathbf{k}}\times\hat{\mathbf{i}}=\hat{\mathbf{j}}, \tag{7b}$$

$$\hat{\mathbf{j}}\times\hat{\mathbf{i}}=-\hat{\mathbf{k}},\qquad \hat{\mathbf{k}}\times\hat{\mathbf{j}}=-\hat{\mathbf{i}},\qquad \hat{\mathbf{i}}\times\hat{\mathbf{k}}=-\hat{\mathbf{j}}. \tag{7c}$$



(+)
$\cdots\ \hat{\mathbf{i}}\ \hat{\mathbf{j}}\ \hat{\mathbf{k}}\ \hat{\mathbf{i}}\ \hat{\mathbf{j}}\ \hat{\mathbf{k}}\ \hat{\mathbf{i}}\ \hat{\mathbf{j}}\ \hat{\mathbf{k}}\ \cdots$
(−)

Figure 3. Mnemonic device for (7b) and (7c).

Equations (7b) and (7c) can be remembered easily if one keeps in mind the mnemonic device shown in Fig. 3. Note carefully, however, that the signs in (7b), (7c), and Fig. 3, are correct only if our x, y, z coordinate system is *right-handed*. If we were to choose a *left-handed* system, the signs in (7b), (7c), and Fig. 3 would be reversed; for example, we would have $\hat{\mathbf{i}}\times\hat{\mathbf{j}}=-\hat{\mathbf{k}}$, and $\hat{\mathbf{j}}\times\hat{\mathbf{k}}=-\hat{\mathbf{i}}$. *As long as we are consistent* it does not matter whether we choose a right-handed system or a left-handed one. But to minimize confusion, we shall *always* adopt right-handed systems, as is standard mathematical practice.

Analogous to the expression (2) for 2-space, any given vector $\mathbf{u}$ in 3-space, say from $(0,0,0)$ to (a,b,c) [or, equivalently, from (x_0, y_0, z_0) to (x_0+a, y_0+b, z_0+c)], can be expressed as a unique linear combination of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, namely,

$$\mathbf{u}=a\hat{\mathbf{i}}+b\hat{\mathbf{j}}+c\hat{\mathbf{k}} \tag{8}$$

as illustrated in Fig. 2. Here a, b, c are called the x, y, z **components** of $\mathbf{u}$, respectively. Instead of the letters a, b, c, it is more common to use the letters u_1, u_2, u_3 for the components of $\mathbf{u}$ so that

$$\mathbf{u}=u_1\hat{\mathbf{i}}+u_2\hat{\mathbf{j}}+u_3\hat{\mathbf{k}}. \tag{9}$$

Since each vector in 3-space can be expressed, or *expanded*, as a unique linear combination of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, we say that the set $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right\}$ is a **basis** for 3-space, an orthonormal basis, in fact. (Bases and expansions are covered in more detail in Section 9.9, but that section is not a prerequisite for the present one.)

The form (9) is especially convenient for calculations. For example, if

$$\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}} \qquad \text{and} \qquad \mathbf{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}, \tag{10}$$

then

$$\begin{aligned}
\mathbf{u}+\mathbf{v} &= (u_1+v_1)\hat{\mathbf{i}} + (u_2+v_2)\hat{\mathbf{j}} + (u_3+v_3)\hat{\mathbf{k}}, && \text{(11a)}\\
\mathbf{u}-\mathbf{v} &= (u_1-v_1)\hat{\mathbf{i}} + (u_2-v_2)\hat{\mathbf{j}} + (u_3-v_3)\hat{\mathbf{k}}, && \text{(11b)}\\
\alpha\mathbf{u} &= (\alpha u_1)\hat{\mathbf{i}} + (\alpha u_2)\hat{\mathbf{j}} + (\alpha u_3)\hat{\mathbf{k}}, && \text{(11c)}\\
\mathbf{u}\cdot\mathbf{v} &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}})\cdot(v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}})\\
&= u_1v_1\hat{\mathbf{i}}\cdot\hat{\mathbf{i}} + u_1v_2\hat{\mathbf{i}}\cdot\hat{\mathbf{j}} + \cdots + u_3v_2\hat{\mathbf{k}}\cdot\hat{\mathbf{j}} + u_3v_3\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}\\
&= u_1v_1 + u_2v_2 + u_3v_3, && \text{(11d)}\\
\mathbf{u}\times\mathbf{v} &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}})\times(v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}})\\
&= u_1v_1\hat{\mathbf{i}}\times\hat{\mathbf{i}} + u_1v_2\hat{\mathbf{i}}\times\hat{\mathbf{j}} + \cdots + u_3v_2\hat{\mathbf{k}}\times\hat{\mathbf{j}} + u_3v_3\hat{\mathbf{k}}\times\hat{\mathbf{k}}\\
&= (u_2v_3-u_3v_2)\hat{\mathbf{i}} - (u_1v_3-u_3v_1)\hat{\mathbf{j}} + (u_1v_2-u_2v_1)\hat{\mathbf{k}}. && \text{(11e)}
\end{aligned}$$

The right-hand side of (11e) happens to be expressible as a third-order determinant so that

$$\boxed{\mathbf{u}\times\mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},} \tag{12}$$

this form being easier to remember than (11e).

EXAMPLE 1. Given

$$\mathbf{u} = \hat{\mathbf{i}} - 2\hat{\mathbf{k}}, \qquad \mathbf{v} = 3\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}, \qquad \mathbf{w} = \hat{\mathbf{j}} - \hat{\mathbf{k}},$$

compute $(\mathbf{u}+2\mathbf{v})\times\mathbf{w}$ and $\mathbf{u}\cdot\mathbf{w}$.

$$\begin{aligned}
(\mathbf{u}+2\mathbf{v})\times\mathbf{w} &= [(1+6)\hat{\mathbf{i}} + (0+2)\hat{\mathbf{j}} + (-2+2)\hat{\mathbf{k}}]\times(\hat{\mathbf{j}}-\hat{\mathbf{k}})\\
&= (7\hat{\mathbf{i}} + 2\hat{\mathbf{j}})\times(\hat{\mathbf{j}}-\hat{\mathbf{k}}) = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 7 & 2 & 0 \\ 0 & 1 & -1 \end{vmatrix} = -2\hat{\mathbf{i}} + 7\hat{\mathbf{j}} + 7\hat{\mathbf{k}}
\end{aligned}$$

and $\mathbf{u}\cdot\mathbf{w} = (1)(0) + (0)(1) + (-2)(-1) = 2.$ ▮

Closure. In this section we introduce the (always right-handed) Cartesian coordinate system x, y, z, and its corresponding orthonormal base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. Non-Cartesian systems are the subject of Section 14.6.

EXERCISES 14.3

1. Let $\mathbf{u} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} - 3\hat{\mathbf{k}}$, $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathbf{w} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{k}}$, $\mathbf{x} = 8\hat{\mathbf{i}} - \hat{\mathbf{j}} - 11\hat{\mathbf{k}}$.

(a) Compute $2\mathbf{u} - \mathbf{v}$, $\mathbf{u}\cdot\mathbf{v}$, $\mathbf{v}\cdot\mathbf{u}$, $\mathbf{u}\times\mathbf{v}$, $\mathbf{v}\times\mathbf{u}$ and $\|\mathbf{u}\times\mathbf{v}\|$.
(b) Compute $(\mathbf{u}+3\mathbf{w})\times\mathbf{v}$, $\|(\mathbf{u}+3\mathbf{w})\times\mathbf{v}\|$, $\mathbf{u}\cdot\mathbf{w}$ and $|\mathbf{u}\cdot\mathbf{w}|$.
(c) Compute $(\mathbf{u}\cdot 2\mathbf{w})\mathbf{v}$, $\mathbf{x}\times(-2\mathbf{u}) + 3\mathbf{v}$, $(\mathbf{u}+\mathbf{w})\times\mathbf{v}$ and $\|\mathbf{x}\|$.
(d) Compute $\mathbf{u} + 2\mathbf{v} - \mathbf{w} + 3\mathbf{x}$, $(2\mathbf{u}+\mathbf{w})\cdot(\mathbf{v}-\mathbf{x})$ and $\mathbf{w}\times(\mathbf{v}-\mathbf{x})$.
(e) Compute $(3\mathbf{v}+\mathbf{w}-\mathbf{x})\cdot\mathbf{v}$, $(\mathbf{w}-\mathbf{x})\times(2\mathbf{v})$ and $\|\mathbf{u}\times(\mathbf{v}\times\mathbf{w})\|$.
(f) With $\mathbf{u}, \mathbf{v}, \mathbf{x}$ tail to tail, show that the three vectors lie in a plane. Do this two ways: first, by taking suitable cross and dot products and, second, by showing that one of the vectors can be expressed as a linear combination of the others.
(g) With $\mathbf{u}, \mathbf{v}, \mathbf{w}$ tail to tail, show — by taking suitable cross and dot products — that the three vectors do *not* lie in the same plane.
(h) Find a vector perpendicular to $\mathbf{u}$ and $\mathbf{v}$. Verify your answer.
(i) Find a vector perpendicular to $\mathbf{u}$ and $\mathbf{w}$. Verify your answer.
(j) Find a vector perpendicular to $\mathbf{v}$ and $\mathbf{w}$. Verify your answer.
(k) With $\mathbf{u}, \mathbf{v}, \mathbf{w}$ tail to tail, find the area of the triangle formed by their heads, and find a vector perpendicular to that triangle..
(l) Repeat part (k) for $\mathbf{u}, \mathbf{v}, \mathbf{x}$.
(m) Repeat part (k) for $\mathbf{u}, \mathbf{w}, \mathbf{x}$.
(n) With $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}$ tail to tail, show whether or not their heads lie in a common plane.

2. Consider three forces: $\mathbf{F}_1 = 2\hat{\mathbf{i}}+\hat{\mathbf{k}}$ with its tail at $(2, 1, -1)$, $\mathbf{F}_2 = \hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}$ with its tail at $(0, 0, 0)$, and $\mathbf{F}_3 = \hat{\mathbf{i}} + 4\hat{\mathbf{j}}$ with its tail at $(1, 1, 1)$.

(a) Find the total moment $\mathbf{M}$ about $(0, 0, 0)$.
(b) Find the total moment $\mathbf{M}$ about $(1, 1, 0)$.
(c) Find the total moment $\mathbf{M}$ about $(8, 6, -10)$.
(d) Find the total moment $\mathbf{M}$ about $(3, 2, 1)$.
(e) Determine, if possible, an additional force $\mathbf{F}_4$ with its tail at $(2, -1, -1)$, such that the total moment $\mathbf{M}$ about $(1, 1, 1)$ is zero. If it is *not* possible, explain that circumstance in geometrical terms.
(f) Same as part (e) with $(2, -1, -1)$ changed to $(3, 6, -1)$.
(g) Same as part (e) with $(2, -1, -1)$ changed to $(-1, -4, 3)$.

3. Use a suitable cross product to determine whether or not the following points lie on a straight line.

(a) $(2, 3, 1)$, $(1, -1, 4)$, $(2, 2, 1)$
(b) $(1, 3, 0)$, $(2, -1, 1)$, $(3, -5, 2)$
(c) $(1, 0, 2)$, $(-4, 3, 0)$, $(-9, 6, -2)$
(d) $(1, 0, 2)$, $(-4, 3, 0)$, $(12, -6, 6)$
(e) $(1, 0, 2)$, $(-4, 3, 0)$, $(-3, 3, 2)$
(f) $(1, 0, 1)$, $(2, -1, 4)$, $(5, 7, -3)$

4. Use a suitable property of the cross product to find the area of the triangles with the following vertices.

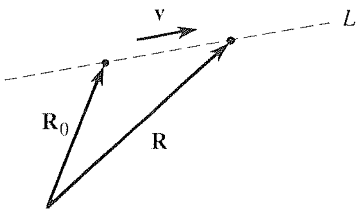
(a) $(1, 4, 3)$, $(2, 0, -1)$, $(0, 0, 5)$
(b) $(2, -2, 1)$, $(4, 0, 3)$, $(2, 3, 5)$
(c) $(0, 0, 2)$, $(0, 4, 0)$, $(3, 1, 1)$
(d) $(8, 7, 4)$, $(2, -5, 0)$, $(1, 6, 8)$
(e) $(1, 1, 1)$, $(1, 1, 2)$, $(3, 2, 1)$
(f) $(5, 4, 1)$, $(3, 2, 3)$, $(-4, 0, 6)$
(g) $(2, -1, 1)$, $(0, 2, 3)$, $(9, 6, 1)$
(h) $(14, 1, -6)$, $(2, 2, 1)$, $(0, 4, 9)$

5. (a) Determine a unit normal vector to the plane $x+2y-z = 5$ by finding three distinct (noncollinear) points A, B, C in the plane, crossing $\mathbf{AB}$ with $\mathbf{AC}$, and normalizing.
(b) Repeat part (a) for the plane $2x + y - z = 0$.
(c) Repeat part (a) for the plane $x - 3y + 4z = -2$.
(d) Repeat part (a) for the plane $x + 3y + z = 4$.
(e) Repeat part (a) for the plane $ax + by + cz = d$.

6. (*Equation of a line*) (a) Let $\mathbf{R}_0$ be a given point in 3-space (i.e., the point at the head of the position vector $\mathbf{R}_0$ issuing from the origin). And let $\mathbf{v}$ be a given (nonzero) vector (see the figure). Show that the locus of points $\mathbf{R}$ which constitute a straight line L through $\mathbf{R}_0$, parallel to $\mathbf{v}$, is determined by either of the equations

$$(\mathbf{R} - \mathbf{R}_0) \times \mathbf{v} = \mathbf{0}, \qquad \text{or} \tag{6.1}$$

$$\mathbf{R} = \mathbf{R}_0 + \mathbf{v}t \qquad (-\infty < t < \infty). \tag{6.2}$$



(b) Using (6.2), show that the straight line through $(2, 5, -1)$, parallel to the vector $4\hat{\mathbf{i}} - \hat{\mathbf{k}}$, may be defined by the parametric equations

$$x = 2 + 4t, \qquad y = 5, \qquad z = -1 - t$$

for $-\infty < t < \infty$.
(c) Determine parametric equations for the straight line through $(1,0,5)$ which is w parallel to the vector $-3\hat{\mathbf{i}}+\hat{\mathbf{j}}+2\hat{\mathbf{k}}$.

7. (*Equation of a plane*) Let $\mathbf{R}_0$ be a position vector from the origin to a given plane, and let $\mathbf{R}_0$ be normal to the plane so that $\|\mathbf{R}_0\|$ is the shortest distance from the origin to the plane. If $\mathbf{R} = x\hat{\mathbf{i}}+y\hat{\mathbf{j}}+z\hat{\mathbf{k}}$ is the position vector to any point (x,y,z) in the plane, and $\hat{\mathbf{n}}$ is a unit normal vector to the plane, then surely $(\mathbf{R}-\mathbf{R}_0)\cdot\hat{\mathbf{n}} = 0$ or, since $\mathbf{R}_0\cdot\hat{\mathbf{n}} = \|\mathbf{R}_0\|$,

$$\mathbf{R}\cdot\hat{\mathbf{n}} = \|\mathbf{R}_0\|\,. \qquad (7.1)$$

Equation (7.1) is the general equation of a plane in vector form and is interesting because each quantity has a clear geometrical significance: $\mathbf{R} = x\hat{\mathbf{i}}+y\hat{\mathbf{j}}+z\hat{\mathbf{k}}$ is a position vector to any point (x,y,z) in the plane, $\hat{\mathbf{n}}$ is a unit vector to the plane (and is unique to within a factor of ± 1), and $\|\mathbf{R}_0\|$ is the shortest distance from the origin to the plane.

(a) Find the shortest distance from the origin to the plane, and a unit vector normal to the plane, for the plane defined by the equation $2x+y-3z=4$.
(b) Repeat (a) for $x-y-z=5$.
(c) Repeat (a) for $3x+y-z=0$.
(d) Repeat (a) for $x+2y+3z=8$.
(e) Repeat (a) for $3y-z=5$.
(f) Repeat (a) for $ax+by+cz=d$.
(g) Repeat (a) for $a(x-x_0)+b(y-y_0)+c(z-z_0)=0$.

8. The vectors $\mathbf{u} = (1,1,2)$ and $\mathbf{v} = (3,2,-1)$ determine a plane (by their span). Find a nonzero vector in that plane that is perpendicular to $\mathbf{w} = (2,4,3)$. As usual, explain your reasoning.

9. Same as Exercise 8, for

(a) $\mathbf{u} = (4,-1,1)$, $\mathbf{v} = (1,1,2)$, $\mathbf{w} = (3,0,5)$
(b) $\mathbf{u} = (1,2,3)$, $\mathbf{v} = (3,2,1)$, $\mathbf{w} = (1,2,4)$
(c) $\mathbf{u} = (1,2,3)$, $\mathbf{v} = (1,0,2)$, $\mathbf{w} = (3,0,5)$
(d) $\mathbf{u} = (1,0,1)$, $\mathbf{v} = (0,1,1)$, $\mathbf{w} = (0,0,1)$

14.4 Multiple Products

We have discussed the dot and cross products of two vectors, $\mathbf{u}\cdot\mathbf{v}$ and $\mathbf{u}\times\mathbf{v}$, respectively. Products of more than two vectors are also encountered and are the subject of this section.

What products of *three* vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are possible? Writing down such combinations, consider

$$\mathbf{u}\cdot(\mathbf{v}\cdot\mathbf{w}), \quad \mathbf{u}\times(\mathbf{v}\cdot\mathbf{w}), \quad \mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}), \quad \mathbf{u}\times(\mathbf{v}\times\mathbf{w}). \qquad (1)$$

Now, $\mathbf{v}\cdot\mathbf{w}$ is a scalar so the first two members of (1) amount to a vector dotted with a scalar and a vector crossed with a scalar, respectively. Such quantities are *not defined* so the first two items in (1) can be discarded. It remains to consider

$$\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w}), \qquad (2)$$

called the **scalar triple product** because the result is a scalar, and

$$\mathbf{u}\times(\mathbf{v}\times\mathbf{w}), \qquad (3)$$

called the **vector triple product** because the result is a vector.

14.4.1. Scalar triple product. First, the scalar triple product $\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})$.* For notational simplicity, we can drop the parentheses and write $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ since $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$

*The triple product $\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})$ is sometimes denoted as "$(\mathbf{uvw})$." That notation is not used in this book.

can mean only $(\mathbf{u}\cdot\mathbf{v})\times\mathbf{w}$ or $\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})$, and of these two the former is not defined and is not a viable candidate.

To observe an important fact about $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$, let us sketch the "$\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped" shown in Fig. 1 (for the case where $\mathbf{u},\mathbf{v},\mathbf{w}$ do not lie in the same plane). Now, $\mathbf{v}\times\mathbf{w}$ is the magnitude (norm) of $\mathbf{v}\times\mathbf{w}$ times the unit vector $\hat{\mathbf{e}}$ shown in Fig. 1. Thus,

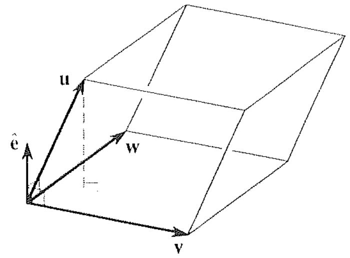


Figure 1. The $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped.

$$\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{u}\cdot\|\mathbf{v}\times\mathbf{w}\|\,\hat{\mathbf{e}}=\|\mathbf{v}\times\mathbf{w}\|\,(\mathbf{u}\cdot\hat{\mathbf{e}}). \tag{4}$$

But $\|\mathbf{v}\times\mathbf{w}\|$ is, according to (10) in Section 14.2, the area of the $\mathbf{v},\mathbf{w}$ parallelogram, that is, the *base area* of the $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped. And $\mathbf{u}\cdot\hat{\mathbf{e}}$ is equal to the *altitude* (the dashed line). Hence it follows from (4) that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ is the volume of the $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped. However, if $\mathbf{u}$ were directed such that $\mathbf{u}\cdot\hat{\mathbf{e}}$ were negative, then $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ would be the *negative* of the volume. In any case, if we introduce absolute values, then

$$\boxed{|\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}|=\text{volume of }\mathbf{u},\mathbf{v},\mathbf{w}\text{ parallelepiped.}} \tag{5}$$

Finally, if we recall from Section 14.3 that

$$\mathbf{v}\times\mathbf{w}=\begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix},$$

where $\mathbf{v}=v_1\hat{\mathbf{i}}+v_2\hat{\mathbf{j}}+v_3\hat{\mathbf{k}}$ and $\mathbf{w}=w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}}+w_3\hat{\mathbf{k}}$, it should not be hard to see that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ can be expressed neatly as

$$\boxed{\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\begin{vmatrix}u_1 & u_2 & u_3\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix}.} \tag{6}$$

Now, interchanging two rows of a determinant changes the sign of the determinant. If, in (6), we interchange the first row with the second, and then the third with the first, the new determinant can be identified [according to the pattern in (6)] as $\mathbf{w}\cdot\mathbf{u}\times\mathbf{v}$. Since there were two sign changes, we see that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{w}\cdot\mathbf{u}\times\mathbf{v}$, or

$$\boxed{\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{u}\times\mathbf{v}\cdot\mathbf{w},} \tag{7}$$

that is, interchanging the dot and the cross leaves the scalar triple product unchanged.

14.4.2. Vector triple product. Next, consider the vector triple product $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$. Here we *do* need to retain the parentheses since $\mathbf{u}\times\mathbf{v}\times\mathbf{w}$ could mean either $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$ or $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$. Both of these quantities are indeed defined, and they are not necessarily equal to each other; for example, $\hat{\mathbf{i}}\times(\hat{\mathbf{i}}\times\hat{\mathbf{j}})=\hat{\mathbf{i}}\times\hat{\mathbf{k}}=-\hat{\mathbf{j}}$, whereas $(\hat{\mathbf{i}}\times\hat{\mathbf{i}})\times\hat{\mathbf{j}}=\mathbf{0}\times\hat{\mathbf{j}}=\mathbf{0}$.

Thus, let us state, for reference, that in general

$$(\mathbf{u}\times\mathbf{v})\times\mathbf{w}\neq\mathbf{u}\times(\mathbf{v}\times\mathbf{w}). \tag{8}$$

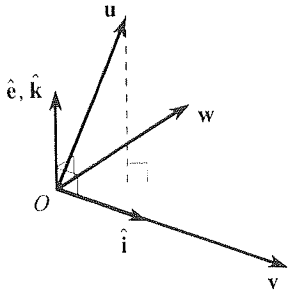


Figure 2. Derivation of (9).

Recall that $\mathbf{v}\times\mathbf{w}$ is perpendicular to the $\mathbf{v}$, $\mathbf{w}$ plane so it must be some multiple of $\hat{\mathbf{e}}$ (Fig. 2). Similarly, $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$ is perpendicular to both $\mathbf{u}$ and $\mathbf{v}\times\mathbf{w}$, hence, to both $\mathbf{u}$ and $\hat{\mathbf{e}}$. Being perpendicular to $\hat{\mathbf{e}}$, it must lie in the $\mathbf{v}$, $\mathbf{w}$ plane so it must be expressible in the form $\alpha\mathbf{v}+\beta\mathbf{w}$. To determine α and β, let us introduce a convenient reference coordinate system and set of base vectors. Specifically, let the point O in Fig. 2 be the origin of a Cartesian coordinate system with the usual base vectors $\hat{\mathbf{i}},\hat{\mathbf{j}},\hat{\mathbf{k}}$. And let the system be oriented so that $\hat{\mathbf{i}}$ is aligned with $\mathbf{v}$, and $\hat{\mathbf{k}}$ coincides with the $\hat{\mathbf{e}}$ shown in Fig. 2. Then $\hat{\mathbf{j}}$ must lie in the $\mathbf{v},\mathbf{w}$ plane so we can express

$$\mathbf{v}=v_1\hat{\mathbf{i}},\qquad \mathbf{w}=w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}},\qquad \mathbf{u}=u_1\hat{\mathbf{i}}+u_2\hat{\mathbf{j}}+u_3\hat{\mathbf{k}}.$$

Hence $\mathbf{v}\times\mathbf{w}=v_1w_2\hat{\mathbf{k}}$, and

$$\begin{aligned}\mathbf{u}\times(\mathbf{v}\times\mathbf{w})&=(u_1\hat{\mathbf{i}}+u_2\hat{\mathbf{j}}+u_3\hat{\mathbf{k}})\times(v_1w_2\hat{\mathbf{k}})\\&=-u_1v_1w_2\hat{\mathbf{j}}+u_2v_1w_2\hat{\mathbf{i}}\\&=(u_1w_1+u_2w_2)v_1\hat{\mathbf{i}}-u_1v_1(w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}})\\&=(u_1w_1+u_2w_2)\mathbf{v}-u_1v_1\mathbf{w}.\end{aligned}$$

But $u_1w_1+u_2w_2=\mathbf{u}\cdot\mathbf{w}$ and $u_1v_1=\mathbf{u}\cdot\mathbf{v}$ so we have the useful formula

$$\boxed{\mathbf{u}\times(\mathbf{v}\times\mathbf{w})=(\mathbf{u}\cdot\mathbf{w})\mathbf{v}-(\mathbf{u}\cdot\mathbf{v})\mathbf{w}.} \tag{9}$$

Closure. We could consider quadruple products such as $\mathbf{u}\times(\mathbf{v}\times(\mathbf{w}\times\mathbf{x}))$ and $(\mathbf{u}\times\mathbf{v})\cdot(\mathbf{w}\times\mathbf{x})$, quintuple products, and so on, but the triple products discussed here will suffice for our purposes. In the subsequent sections, the scalar triple product will occur more than the vector triple product so keep in mind the geometrical property (5), the computational formula (6), and the property (7).

EXERCISES 14.4

1. If $\mathbf{u}=\hat{\mathbf{i}}-\hat{\mathbf{j}}$, $\mathbf{v}=2\hat{\mathbf{i}}+\hat{\mathbf{j}}+3\hat{\mathbf{k}}$, $\mathbf{w}=\hat{\mathbf{j}}-\hat{\mathbf{k}}$, evaluate

(a) $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ and $\mathbf{u}\times\mathbf{v}\cdot\mathbf{w}$
(b) $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$ and $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$

2. If $\mathbf{A}=\hat{\mathbf{i}}+\hat{\mathbf{j}}-\hat{\mathbf{k}}$, $\mathbf{B}=3\hat{\mathbf{j}}-\hat{\mathbf{k}}$, $\mathbf{C}=\hat{\mathbf{i}}+2\hat{\mathbf{k}}$, evaluate

(a) $\mathbf{A}\cdot\mathbf{B}\times\mathbf{C}$ and $\mathbf{A}\times\mathbf{B}\cdot\mathbf{C}$
(b) $\mathbf{A}\times(\mathbf{B}\times\mathbf{C})$ and $(\mathbf{A}\times\mathbf{B})\times\mathbf{C}$

3. Verify the identity (9) by working out the left- and right-hand sides separately, for each case.

(a) $\mathbf{u}=\hat{\mathbf{i}}$, $\mathbf{v}=\hat{\mathbf{j}}$, $\mathbf{w}=\hat{\mathbf{k}}$
(b) $\mathbf{u}=\hat{\mathbf{i}}$, $\mathbf{v}=\hat{\mathbf{i}}$, $\mathbf{w}=\hat{\mathbf{j}}$
(c) $\mathbf{u}=\hat{\mathbf{j}}-\hat{\mathbf{k}}$, $\mathbf{v}=\hat{\mathbf{i}}+3\hat{\mathbf{j}}$, $\mathbf{w}=\hat{\mathbf{i}}+2\hat{\mathbf{k}}$
(d) $\mathbf{u}=8\hat{\mathbf{i}}+6\hat{\mathbf{j}}$, $\mathbf{v}=\hat{\mathbf{i}}-4\hat{\mathbf{k}}$, $\mathbf{w}=2\hat{\mathbf{j}}$

4. Find the volume of the parallelepiped having the following vectors as adjacent edges.

(a) $\hat{\mathbf{i}}-2\hat{\mathbf{j}}$, $3\hat{\mathbf{j}}$, $\hat{\mathbf{i}}+\hat{\mathbf{k}}$ (b) $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}$

(c) $\hat{\mathbf{i}}+2\hat{\mathbf{j}},\ 3\hat{\mathbf{i}},\ \hat{\mathbf{j}}+\hat{\mathbf{k}}$ (d) $3\hat{\mathbf{j}},\ 2\hat{\mathbf{i}}-\hat{\mathbf{j}},\ \hat{\mathbf{i}}+5\hat{\mathbf{k}}$
(e) $\hat{\mathbf{i}},\ \hat{\mathbf{i}}+\hat{\mathbf{j}},\ \hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}$ (f) $\hat{\mathbf{i}}+\hat{\mathbf{j}},\ \hat{\mathbf{i}},\ \hat{\mathbf{i}}+\hat{\mathbf{j}}+\hat{\mathbf{k}}$

5. Show that $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}=(\mathbf{w}\cdot\mathbf{u})\mathbf{v}-(\mathbf{w}\cdot\mathbf{v})\mathbf{u}$. You may use (9).

6. Show that $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})+\mathbf{w}\times(\mathbf{u}\times\mathbf{v})+\mathbf{v}\times(\mathbf{w}\times\mathbf{u})=\mathbf{0}$.

7. If $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, are distinct vectors issuing from a common point, show that the vector $(\mathbf{A}\times\mathbf{B})+(\mathbf{B}\times\mathbf{C})+(\mathbf{C}\times\mathbf{A})$ is perpendicular to the plane containing their heads.

8. (*Orthogonal separation*) Let a given (nonzero) vector $\mathbf{u}$ be separated into the sum of two orthogonal vectors, one parallel to a given (nonzero) vector $\mathbf{v}$ and the other perpendicular to $\mathbf{v}$ (i.e., $\mathbf{u}=\mathbf{u}_{\text{par}}+\mathbf{u}_{\text{perp}}$). Show that

$$\mathbf{u}_{\text{par}}=(\mathbf{u}\cdot\hat{\mathbf{v}})\hat{\mathbf{v}},\qquad \mathbf{u}_{\text{perp}}=\hat{\mathbf{v}}\times(\mathbf{u}\times\hat{\mathbf{v}}),\tag{8.1}$$

where $\hat{\mathbf{v}}=\mathbf{v}/\|\mathbf{v}\|$.

9. Prove the following identities involving quadruple products. HINT: Use (7) and (9).

(a) $(\mathbf{A}\times\mathbf{B})\cdot(\mathbf{C}\times\mathbf{D})=(\mathbf{A}\cdot\mathbf{C})(\mathbf{B}\cdot\mathbf{D})-(\mathbf{A}\cdot\mathbf{D})(\mathbf{B}\cdot\mathbf{C})$
This is **Lagrange's identity**, of which equation (5.1) of Exercise 5 in Section 14.2 is a special case.
(b) $(\mathbf{A}\times\mathbf{B})\times(\mathbf{C}\times\mathbf{D})=(\mathbf{A}\cdot\mathbf{B}\times\mathbf{D})\mathbf{C}-(\mathbf{A}\cdot\mathbf{B}\times\mathbf{C})\mathbf{D}$

10. In deriving (9), we introduced the orthonormal basis $\{\hat{\mathbf{i}},\hat{\mathbf{j}},\hat{\mathbf{k}}\}$, oriented in such a way that $\mathbf{v}=v_1\hat{\mathbf{i}}$, $\mathbf{w}=w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}}$, and $\mathbf{u}=u_1\hat{\mathbf{i}}+u_2\hat{\mathbf{j}}+u_3\hat{\mathbf{k}}$. Although this orientation was *convenient* (since then $\mathbf{v}$ had only one component, and $\mathbf{w}$ had only two, so that the calculations were rendered as short as possible), we claim that it was not essential. Verify this claim by rederiving (9) using the expansions $\mathbf{u}=u_1\hat{\mathbf{i}}+u_2\hat{\mathbf{j}}+u_3\hat{\mathbf{k}}$, $\mathbf{v}=v_1\hat{\mathbf{i}}+v_2\hat{\mathbf{j}}+v_3\hat{\mathbf{k}}$, $\mathbf{w}=w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}}+w_3\hat{\mathbf{k}}$, which, of course, hold for *any* orientation.

11. (*Linear dependence of* $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$) Prove that vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are linearly dependent if and only if $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=0$.

12. Do the following sets of four points lie in a plane? HINT: You may use the result stated in Exercise 11, without proving it, if you wish.

(a) $(1,3,0)$, $(2,1,-1)$, $(0,0,4)$, $(5,0,8)$
(b) $(2,1,-1)$, $(1,3,0)$, $(5,0,9)$, $(0,0,4)$
(c) $(4,0,0)$, $(0,1,0)$, $(1,2,-4)$, $(0,0,1)$
(d) $(0,0,0)$, $(2,6,-1)$, $(1,0,1)$, $(1,2,0)$
(e) $(1,0,1)$, $(2,1,3)$, $(1,-1,0)$, $(3,-1,2)$
(f) $(0,0,2)$, $(0,1,3)$, $(1,2,3)$, $(2,3,4)$

13. Give an example of three vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ (*not* the ones given in the text) such that $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}\neq\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$, and an example of three vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ such that $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}=\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$. (The latter part of the exercise is simple if we let at least one of the three vectors be $\mathbf{0}$. To make the problem more interesting, let us insist that $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ all be nonzero.)

14.5 Differentiation of a Vector Function of a Single Variable

Suppose a particle moves about in 3-space so that its position vector $\mathbf{R}(t)$, from the origin to the particle, is a function of the time t. By definition, its velocity vector $\mathbf{v}(t)$ is the derivative $\mathbf{v}=d\mathbf{R}/dt$. This is but one example of many, where we need to differentiate a vector function. Although we discuss vectors at length in Chapter 9 and in the Sections 14.1–14.4, we have not yet introduced the concept of the differentiataion of a vector function. In fact, we have not yet introduced the concept of a vector function!

Suppose a vector $\mathbf{u}(\tau)$ is defined for each value of a real variable τ on some τ interval. Then we say that $\mathbf{u}$ is a **vector function** of τ on that interval. The physical nature of $\mathbf{u}$, and of τ, is in no way restricted. For instance, $\mathbf{u}$ might represent the position vector of a particle or an electric field vector, and its argument τ might be the time, arc length along a curve, or it might be a purely mathematical parameter.

To define the derivative $d\mathbf{u}/d\tau$, or $\mathbf{u}'(\tau)$, we stay as close as possible to the definition of the derivative given in the calculus, and define

$$\boxed{\frac{d\mathbf{u}}{d\tau} = \mathbf{u}'(\tau) \equiv \lim_{\Delta\tau\to 0} \frac{\mathbf{u}(\tau+\Delta\tau)-\mathbf{u}(\tau)}{\Delta\tau}.} \tag{1}$$

If the limit on the right-hand side exists, for a given value of τ, we say that $\mathbf{u}(\tau)$ is **differentiable** at that τ, and that $\mathbf{u}'(\tau)$ is its **derivative**.

Observe that the right-hand side of (1) is of the general form

$$\lim_{\tau\to\tau_0} \mathbf{F}(\tau). \tag{2}$$

Thus, our definition (1) is meaningful only if we also define the **limit of a vector sequence**, which we do as follows. By

$$\lim_{\tau\to\tau_0} \mathbf{F}(\tau) = \mathbf{L} \tag{3}$$

[i.e., $\mathbf{F}(\tau)$ tends to $\mathbf{L}$ as τ tends to τ_0] we shall mean that

$$\lim_{\tau\to\tau_0} \|\mathbf{F}(\tau)-\mathbf{L}\| = 0, \tag{4}$$

where the limit in (4) is the ordinary limit defined in the calculus since the norm $\|\mathbf{F}(\tau)-\mathbf{L}\|$ is a scalar function of τ. Further, we say that $\mathbf{F}(\tau)$ is **continuous** at τ_0 if $\lim_{\tau\to\tau_0}\mathbf{F}(\tau)=\mathbf{F}(\tau_0)$.

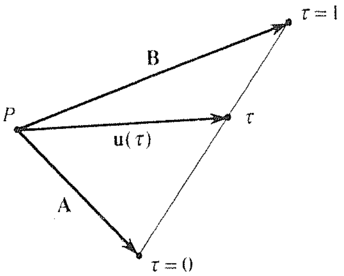


Figure 1. $\mathbf{u}(\tau)$ given by (5).

EXAMPLE 1. To illustrate (1), suppose that

$$\mathbf{u}(\tau) = \mathbf{A} + \tau^3(\mathbf{B}-\mathbf{A}), \qquad (0 \le \tau \le 1) \tag{5}$$

where $\mathbf{A}$ and $\mathbf{B}$ are given vectors that do not vary with τ. If the tail of $\mathbf{u}$ is fixed at P, then $\mathbf{u}(\tau)$ varies with τ as shown in Fig. 1. Then (1) gives

$$\begin{aligned} \mathbf{u}'(\tau) &= \lim_{\Delta\tau\to 0} \frac{\mathbf{A}+(\tau+\Delta\tau)^3(\mathbf{B}-\mathbf{A})-\mathbf{A}-\tau^3(\mathbf{B}-\mathbf{A})}{\Delta\tau} \\ &= \lim_{\Delta\tau\to 0} \frac{3\tau^2\Delta\tau+3\tau(\Delta\tau)^2+(\Delta\tau)^3}{\Delta\tau}(\mathbf{B}-\mathbf{A}) \\ &= \lim_{\Delta\tau\to 0} \left[3\tau^2+3\tau\Delta\tau+(\Delta\tau)^2\right](\mathbf{B}-\mathbf{A}) = 3\tau^2(\mathbf{B}-\mathbf{A}). \end{aligned} \tag{6}$$

To illustrate, we have shown $\mathbf{u}'(\tau)$ at $\tau = 0.5$ in Fig. 2. ∎

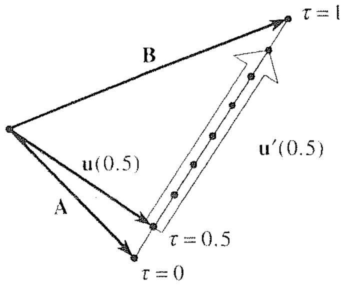


Figure 2. $\mathbf{u}'(\tau)$ at $\tau = 0.5$.

Just as we have rules such as $(uv)' = u'v + uv'$ for the differentiation of scalar functions, we have a number of such rules, which we now list, for the differentiation of vector functions:

$$
\begin{aligned}
(f\mathbf{u})' &= f'\mathbf{u} + f\mathbf{u}', \qquad \text{where } f = f(\tau) && \text{(7a)}\\
(\alpha\mathbf{u} + \beta\mathbf{v})' &= \alpha\mathbf{u}' + \beta\mathbf{v}', \qquad \alpha, \beta \ \text{ constants} && \text{(7b)}\\
(\mathbf{u}\cdot\mathbf{v})' &= \mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}', && \text{(7c)}\\
(\mathbf{u}\times\mathbf{v})' &= \mathbf{u}'\times\mathbf{v} + \mathbf{u}\times\mathbf{v}', && \text{(7d)}\\
(\mathbf{u}\cdot\mathbf{v}\times\mathbf{w})' &= \mathbf{u}'\cdot\mathbf{v}\times\mathbf{w} + \mathbf{u}\cdot\mathbf{v}'\times\mathbf{w} + \mathbf{u}\cdot\mathbf{v}\times\mathbf{w}', && \text{(7e)}\\
[\mathbf{u}\times(\mathbf{v}\times\mathbf{w})]' &= \mathbf{u}'\times(\mathbf{v}\times\mathbf{w}) + \mathbf{u}\times(\mathbf{v}'\times\mathbf{w}) + \mathbf{u}\times(\mathbf{v}\times\mathbf{w}'), && \text{(7f)}\\
\frac{d}{d\tau}\mathbf{u}(f(\tau)) &= \frac{d\mathbf{u}}{df}f'(\tau), \qquad \text{(chain rule)} && \text{(7g)}
\end{aligned}
$$

provided, of course, that the needed derivatives do exist.

As representative, let us prove (7c). Using the definition (1), we find that

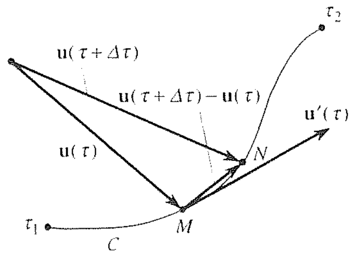
$$
\begin{aligned}
(\mathbf{u}\cdot\mathbf{v})' &= \lim_{\Delta\tau\to 0}\frac{\mathbf{u}(\tau+\Delta\tau)\cdot\mathbf{v}(\tau+\Delta\tau) - \mathbf{u}(\tau)\cdot\mathbf{v}(\tau)}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\frac{\mathbf{u}(\tau+\Delta\tau)\cdot\{\mathbf{v}(\tau) + [\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)]\} - \mathbf{u}(\tau)\cdot\mathbf{v}(\tau)}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\frac{[\mathbf{u}(\tau+\Delta\tau) - \mathbf{u}(\tau)]\cdot\mathbf{v}(\tau) + \mathbf{u}(\tau+\Delta\tau)\cdot[\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)]}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\left[\frac{\mathbf{u}(\tau+\Delta\tau) - \mathbf{u}(\tau)}{\Delta\tau}\right]\cdot\mathbf{v}(\tau) + \lim_{\Delta\tau\to 0}\mathbf{u}(\tau+\Delta\tau)\cdot\left[\frac{\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)}{\Delta\tau}\right]\\
&= \mathbf{u}'(\tau)\cdot\mathbf{v}(\tau) + \mathbf{u}(\tau)\cdot\mathbf{v}'(\tau). \qquad \text{(8)}
\end{aligned}
$$

Proofs of the other formulas are left for the exercises.

To illustrate the use of these rules, let us differentiate the $\mathbf{u}(\tau)$ given by (5) again, but instead of using the cumbersome difference quotient definition (1) let us use the relevant formulas in (7):

$$
\begin{aligned}
\mathbf{u}'(\tau) &= \frac{d}{d\tau}[\mathbf{A} + \tau^3(\mathbf{B} - \mathbf{A})]\\
&= \frac{d}{d\tau}\mathbf{A} + \frac{d}{d\tau}[\tau^3(\mathbf{B} - \mathbf{A})] \qquad \text{per (7b) with } \alpha = \beta = 1\\
&= 0 + \frac{d}{d\tau}(\tau^3)(\mathbf{B} - \mathbf{A}) + \tau^3\frac{d}{d\tau}(\mathbf{B} - \mathbf{A}) \qquad \text{per (7a)}\\
&= 3\tau^2(\mathbf{B} - \mathbf{A}) + \mathbf{0}\\
&= 3\tau^2(\mathbf{B} - \mathbf{A}), \qquad \text{(9)}
\end{aligned}
$$

as found in Example 1.



τ_2
$\mathbf{u}(\tau+\Delta\tau)$
$\mathbf{u}(\tau+\Delta\tau)-\mathbf{u}(\tau)$
$\mathbf{u}'(\tau)$
$\mathbf{u}(\tau)$
N
τ_1
M
C

Figure 3. Graphical interpretation of $\mathbf{u}(\tau)$.

The definition (1) admits a graphical interpretation. Suppose that the tail of the $\mathbf{u}(\tau)$ vector is "bound" at a fixed point P, and that the head of the vector traces a space curve C as τ varies continuously over some interval $\tau_1 \leq \tau \leq \tau_2$, for example, as sketched in Fig. 3. The difference $\mathbf{u}(\tau + \Delta\tau) - \mathbf{u}(\tau)$, in (1) is the vector $\mathbf{MN}$. The endpoint M is fixed, for a given value of τ, and the other endpoint N moves along C and tends to M as $\Delta\tau \to 0$. The limiting direction of

$[\mathbf{u}(\tau+\Delta\tau)-\mathbf{u}(\tau)]/\Delta\tau$ as $\Delta\tau \to 0$, that is, the direction of $\mathbf{u}'(\tau)$ defines a straight line called the **tangent line** to C at τ [provided that $\mathbf{u}(\tau)$ is indeed differentiable there, and that $\mathbf{u}'(\tau) \neq \mathbf{0}$, for the zero vector would not imply any particular direction], and $\mathbf{u}'(\tau)$ is called a **tangent vector**. In Example 1, for instance, C is simply a straight line, and $\mathbf{u}'(\tau)$ was indeed found to be tangent to that line.

EXAMPLE 2. Suppose that a particle is located, with respect to a plane Cartesian coordinate system, by the position vector

$$\mathbf{R}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} = t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}, \tag{10}$$

where t is the time. Then $\mathbf{R}'(t) = \mathbf{v}(t)$ is called the **velocity** of the particle, and $\mathbf{R}''(t) = \mathbf{v}'(t) = \mathbf{a}(t)$ is called its **acceleration**:

$$\begin{aligned} \mathbf{v}(t) = \mathbf{R}'(t) &= \frac{d}{dt}(t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}) \\ &= \frac{d}{dt}(t\hat{\mathbf{i}}) + \frac{d}{dt}(t^2\hat{\mathbf{j}}) \qquad \text{per (7b) with } \alpha = \beta = 1 \\ &= \frac{d}{dt}(t)\hat{\mathbf{i}} + \frac{d}{dt}(t^2)\hat{\mathbf{j}} \qquad \text{per (7a)} \\ &= \hat{\mathbf{i}} + 2t\hat{\mathbf{j}}, \end{aligned} \tag{11}$$

and, similarly,

$$\mathbf{a}(t) = \mathbf{v}'(t) = 2\hat{\mathbf{j}}. \tag{12}$$

In this case the acceleration happens to be a constant vector. The vectors $\mathbf{R}$, $\mathbf{v}$, and $\mathbf{a}$ are shown in Fig. 4 at the instant $t = 1$, at which $\mathbf{R}(1) = \hat{\mathbf{i}}+\hat{\mathbf{j}}$, $\mathbf{v}(1) = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, and $\mathbf{a}(1) = 2\hat{\mathbf{j}}$. Since $x = t$ and $y = t^2$, it follows that $y = x^2$ on C. When $t = 1$ we have $x = 1$, and the slope of C is $dy/dx = 2x = 2$, which is identical to the slope of $v(1) = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, so $\mathbf{v}(1)$ is indeed tangent to C, as stated above this example. Of course, $\mathbf{v}(t)$ is tangent to C for every time t; we chose $t = 1$ just to illustrate. ▮

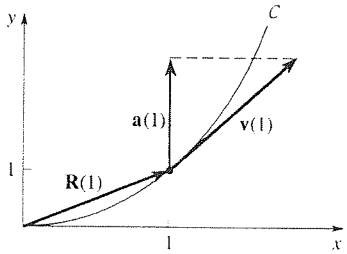


Figure 4. The motion (10).

Closure. We introduce the idea of a vector function, $\mathbf{u}(\tau)$, and define its derivative $\mathbf{u}'(\tau)$ as the limit of the difference quotient given in (1). Because this definition echoes the definition of the derivative of a *scalar* function, familiar from the differential calculus, there are no surprises among the rules listed in (7). Thus, you should feel quite comfortable with the result that if $\mathbf{R}(t) = t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}$, then $\mathbf{R}'(t) = \hat{\mathbf{i}} + 2t\hat{\mathbf{j}}$, given in (11), for instance.

EXERCISES 14.5

1. For each of the following $\mathbf{u}(\tau)$ vectors, determine $\mathbf{u}'$, $\mathbf{u}''$, and $\|\mathbf{u}'\|$.

(a) $\mathbf{A} + \tau^2\mathbf{B}$ ($\mathbf{A}, \mathbf{B}$ constant vectors)

(b) $6\hat{\mathbf{i}} - e^{\tau}\hat{\mathbf{j}}$

(c) $\tau^2\hat{\mathbf{i}} - 4\hat{\mathbf{j}} + 3\cos 2\tau\hat{\mathbf{k}}$

(d) $\cos\tau\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}} + \sin\tau\hat{\mathbf{k}}$

(e) $e^{-\tau}(\cos 2\tau\hat{\mathbf{i}} + \sin 2\tau\hat{\mathbf{j}})$
(f) $e^{\tau}(\cos 2\tau\hat{\mathbf{i}} + \sin 2\tau\hat{\mathbf{j}})$
(g) $\cos 3\tau\hat{\mathbf{i}} - 4\hat{\mathbf{j}} - \sin 3\tau\hat{\mathbf{k}}$
(h) $3\tau\hat{\mathbf{i}} - \hat{\mathbf{j}} + 2\tau^3\hat{\mathbf{k}}$

2. (a) Letting $\mathbf{u} = \tau^2\hat{\mathbf{i}} - \tau\hat{\mathbf{j}} - 3\tau\hat{\mathbf{k}}$, $\mathbf{v} = 2\cos\tau\hat{\mathbf{i}} - \tau\hat{\mathbf{k}}$, and $\mathbf{w} = \sin\tau\hat{\mathbf{j}}$, verify (7c) by working out the left- and right-hand sides.
(b) Repeat part (a), for (7d).
(c) Repeat part (a), for (7e).
(d) Repeat part (a), for (7f).

3. (a) Letting $\mathbf{u} = 3\tau\hat{\mathbf{i}} + \tau^4\hat{\mathbf{k}}$, $\mathbf{v} = \hat{\mathbf{i}} - 4\tau\hat{\mathbf{j}}$, and $\mathbf{w} = 5\tau^2\hat{\mathbf{i}} + \hat{\mathbf{j}} - \tau\hat{\mathbf{k}}$, verify (7c) by working out the left- and right-hand sides.
(b) Repeat part (a), for (7d).
(c) Repeat part (a), for (7e).
(d) Repeat part (a), for (7f).

4. Use the definition (1) to

(a) prove (7a)
(b) prove (7d)

5. (a) Show that (7e) follows from (7c) and (7d).
(b) Show that (7f) follows from (7d).

6. Derive (7g). HINT: Express $\mathbf{u}(f(\tau)) = u_1(f(\tau))\hat{\mathbf{i}} + u_2(f(\tau))\hat{\mathbf{j}} + u_3(f(\tau))\hat{\mathbf{k}}$.

7. (a) Obtain an expression for $(\mathbf{u}\cdot\mathbf{v})''$ analogous to the expression for $(\mathbf{u}\cdot\mathbf{v})'$ given by (7c).
(b) Obtain an expression for $(\mathbf{u}\times\mathbf{v})''$ analogous to the expression for $(\mathbf{u}\times\mathbf{v})'$ given by (7d).

8. (a) Derive the formula

$$\|\mathbf{u}\|' = \frac{\mathbf{u}\cdot\mathbf{u}'}{\|\mathbf{u}\|}. \tag{8.1}$$

(b) Show, from (8.1), that *if* $\|\mathbf{u}(\tau)\|$ = *constant* $\neq 0$, *then* $\mathbf{u}'(\tau)$ *is necessarily either perpendicular to* $\mathbf{u}(\tau)$, *or* $\mathbf{0}$.

9. In (8), we give an intrinsic proof of (7c), that is, one that does not rely on any coordinate system. Alternatively, prove (7c) by introducing Cartesian coordinates so that $\mathbf{u}(\tau) = u_1(\tau)\hat{\mathbf{i}} + u_2(\tau)\hat{\mathbf{j}} + u_3(\tau)\hat{\mathbf{k}}$ and $\mathbf{v}(\tau) = v_1(\tau)\hat{\mathbf{i}} + v_2(\tau)\hat{\mathbf{j}} + v_3(\tau)\hat{\mathbf{k}}$, differentiating $\mathbf{u}(\tau)\cdot\mathbf{v}(\tau)$, and then identifying the result as $\mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}'$.

14.6 Non-Cartesian Coordinates (Optional)

In preceding sections we worked either with no reference coordinate system at all, or with a Cartesian system. Often, it is convenient to work with a coordinate system other than Cartesian, a "non-Cartesian" system. Our purpose in the present section is to introduce three important non-Cartesian systems: plane polar, cylindrical, and spherical. We choose to do this within the context of a physical application, the kinematics of a point (or particle). Specifically, we will develop expressions for the position vector $\mathbf{R}$, the velocity vector $\mathbf{v}$, and the acceleration $\mathbf{a}$ of a point as it moves about in space, in terms of these three coordinate systems.

For reference, let us first note the corresponding results in a Cartesian system x, y, z, with base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. If

$$\mathbf{R}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}} \tag{1}$$

denotes the **position vector** from the origin of a (stationary) Cartesian coordinate system to the point $(x(t), y(t), z(t))$, where t is the time, then the *velocity* $\mathbf{v}$ and *acceleration* $\mathbf{a}$ of the point are

$$\mathbf{v}(t) = \mathbf{R}'(t) = x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}} + z'(t)\hat{\mathbf{k}}, \tag{2}$$

$$\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{R}''(t) = x''(t)\hat{\mathbf{i}} + y''(t)\hat{\mathbf{j}} + z''(t)\hat{\mathbf{k}}, \tag{3}$$

respectively, where primes denote differentiation with respect to the time t.

14.6.1. Plane polar coordinates. The coordinates r, θ and their respective unit base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ are shown in Fig. 1; r, θ are related to the reference x, y coordinates according to

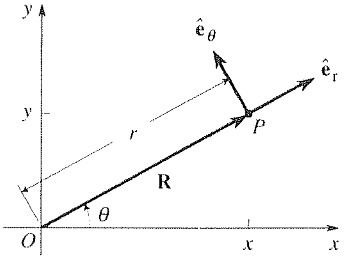


Figure 1. Plane polar coordinates.

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta, \end{aligned} \tag{4a}$$

or, solving (4a) for r and θ,

$$\begin{aligned} r &= \sqrt{x^2+y^2}, \\ \theta &= \tan^{-1}\frac{y}{x}, \end{aligned} \tag{4b}$$

where θ is undefined at the origin. It will be understood that θ is measured in radians unless we specify that it is measured in degrees.

In general, both r and θ may be positive, negative, or zero, but in this book we choose always to have $r \geq 0$ so that r is the *distance* from the origin O to the point in question, say P. In addition, if we limit θ according to $0 \leq \theta < 2\pi$, each point in the x, y plane (except for the origin) corresponds to a unique value of θ.*

The base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ at any point P (other than the origin) are in the positive r, θ directions, respectively. Observe that while $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ are *constant* vectors,† $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ are not! Specifically, $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ vary with θ, although not with r. For example, if we imagine sliding $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ to a new r location, say $r + \Delta r$, with θ held fixed, we see that both vectors remain unchanged because both their magnitude (unity) and their direction remain unchanged, even though their location is different. But if we vary θ, with r fixed, $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ both rotate. The upshot is that

$$\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta), \qquad \hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta). \tag{5}$$

To obtain expressions for the velocity $\mathbf{v}$ and acceleration $\mathbf{a}$ of a point P moving in the x, y plane, we begin with the position vector from O to P (Fig. 1), namely,

$$\boxed{\mathbf{R} = r\hat{\mathbf{e}}_r.} \tag{6}$$

Before proceeding, two points should be made in connection with (6). First, it is tempting to write "$\mathbf{R} = r\hat{\mathbf{e}}_r + \theta\hat{\mathbf{e}}_\theta$," in place of (6), in an effort to follow the pattern

*Unfortunately, the relation $\theta = \tan^{-1} y/x$, given in (4b), does not completely suffice to determine the unique value of θ at a given point without taking into account the signs of x and y. For example, if $x = y = 1$, then surely $\theta = \pi/4$, whereas $\theta = \tan^{-1} 1$ gives both (the correct value) $\theta = \pi/4$ and the (incorrect value) $\theta = 5\pi/4$. Such ambiguity can always be resolved by the original relations $x = r\cos\theta$, $y = r\sin\theta$. For example, for the case $x = y = 1$ mentioned above, $\theta = \pi/4$ satisfies the relations $1 = \sqrt{2}\cos\theta$, $1 = \sqrt{2}\sin\theta$, whereas $\theta = 5\pi/4$ does not and is to be discarded.

†We should qualify that statement: Cartesian base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ do not vary with the *space* coordinates x, y, z. They may, however, vary with *time* if the coordinate system itself is rotating with respect to the observer.

in the Cartesian formula $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$. However, the quoted expression is incorrect; the correct version is given by (6), since (from Fig. 1) the magnitude of $\mathbf{R}$ is r and its direction is $\hat{\mathbf{e}}_r$. Second, is it not strange that θ does not appear in (6)? The answer is that θ *does* appear in (6) since, according to (5), $\hat{\mathbf{e}}_r$ is a function of θ; that is, (6) is really $\mathbf{R} = r\hat{\mathbf{e}}_r(\theta)$.

Now, if the point P moves about according to $r = r(t)$, $\theta = \theta(t)$, then

$$\mathbf{R} = r(t)\hat{\mathbf{e}}_r(\theta(t)), \tag{7}$$

so its velocity is [recall (7a) in Section 14.5]

$$\mathbf{v}(t) = \dot{\mathbf{R}} = \dot{r}\hat{\mathbf{e}}_r + r\dot{\hat{\mathbf{e}}}_r, \tag{8}$$

where we choose to use overhead dots, rather than the more usual primes, to denote differentiation with respect to the time t. Since $\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta(t))$, we can use chain differentiation to express

$$\dot{\hat{\mathbf{e}}}_r = \frac{d}{dt}\hat{\mathbf{e}}_r(\theta(t)) = \frac{d\hat{\mathbf{e}}_r}{d\theta}\frac{d\theta}{dt} = \dot{\theta}\frac{d\hat{\mathbf{e}}_r}{d\theta}. \tag{9}$$

Fine, but what is $d\hat{\mathbf{e}}_r/d\theta$? Since we are working with the base vectors $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$, we wish to expand *all* vectors in terms of these base vectors. Thus, we wish to expand $d\hat{\mathbf{e}}_r/d\theta$ in the form $(\ )\hat{\mathbf{e}}_r + (\ )\hat{\mathbf{e}}_\theta$. To accomplish that, let us fall back on the definition of the derivative,

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \lim_{\Delta\theta\to 0}\frac{\hat{\mathbf{e}}_r(\theta + \Delta\theta) - \hat{\mathbf{e}}_r(\theta)}{\Delta\theta}. \tag{10}$$

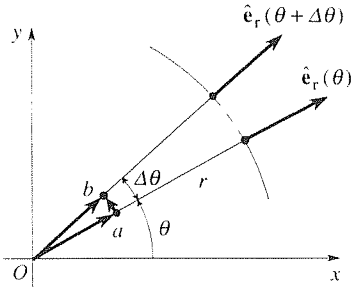

$\hat{\mathbf{e}}_r(\theta + \Delta\theta)$
$\hat{\mathbf{e}}_r(\theta)$
y
b
$\Delta\theta$
r
a
θ
O
x

Figure 2. Calculation of $d\hat{\mathbf{e}}_r/d\theta$.

The vector difference in the numerator is readily evaluated if we slide the two vectors back to O (Fig. 2), so that they are tail-to-tail. Then $\hat{\mathbf{e}}_r(\theta + \Delta\theta) - \hat{\mathbf{e}}_r(\theta)$ is the little vector from a to b; its length is $1\Delta\theta$ and its direction, in the limit as $\Delta\theta \to 0$, is $\hat{\mathbf{e}}_\theta$.* Thus,

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \lim_{\Delta\theta\to 0}\frac{(1\Delta\theta)\hat{\mathbf{e}}_\theta}{\Delta\theta} = \hat{\mathbf{e}}_\theta \tag{11}$$

so (9) becomes $\dot{\hat{\mathbf{e}}}_r = \dot{\theta}\hat{\mathbf{e}}_\theta$, and hence (8) becomes

$$\boxed{\mathbf{v}(t) = \dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta.} \tag{12}$$

The expressions $\dot{r}, r\dot{\theta}$ for the r, θ velocity components, respectively, may well be familiar to you from a course in introductory physics.

Continuing, we differentiate once more to obtain the acceleration,

$$\mathbf{a}(t) = \dot{\mathbf{v}}(t) = \ddot{r}\hat{\mathbf{e}}_r + \dot{r}\dot{\hat{\mathbf{e}}}_r + \dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta + r\ddot{\theta}\hat{\mathbf{e}}_\theta + r\dot{\theta}\dot{\hat{\mathbf{e}}}_\theta. \tag{13}$$

*As $\Delta\theta \to 0$, the length of the vector from a to b, say $\mathbf{ab}$, approaches the arc length of the circular arc through those points and centered at O. Thus, it follows from the formula $s = r\theta$ for the arc length s of a circular arc of radius r and angle θ, that $\|\mathbf{ab}\| \sim 1\Delta\theta$ as $\Delta\theta \to 0$.

We have already seen that $\dot{\hat{e}}_r = \dot{\theta}\hat{e}_\theta$, but we still need to evaluate the $\dot{\hat{e}}_\theta$ in (13). We have

$$\dot{\hat{e}}_\theta = \frac{d}{dt}\hat{e}_\theta[\theta(t)] = \frac{d\hat{e}_\theta}{d\theta}\frac{d\theta}{dt} = \dot{\theta}\frac{d\hat{e}_\theta}{d\theta}, \tag{14}$$

where

$$\frac{d\hat{e}_\theta}{d\theta} = \lim_{\Delta\theta\to 0}\frac{\hat{e}_\theta(\theta+\Delta\theta)-\hat{e}_\theta(\theta)}{\Delta\theta}. \tag{15}$$

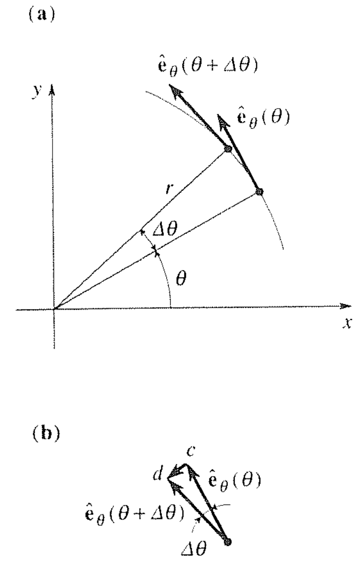


Figure 3. Calculation of $d\hat{e}_\theta/d\theta$.

To evaluate the difference $\hat{e}_\theta(\theta+\Delta\theta)-\hat{e}_\theta(\theta)$, where these two vectors are shown in Fig. 3a, we pick them up and place them tail to tail (without altering their orientation), as shown in Fig. 3b. Their difference is the little vector from c to d, namely, $(1\,\Delta\theta)(-\hat{e}_r)$ so that

$$\frac{d\hat{e}_\theta}{d\theta} = \lim_{\Delta\theta\to 0}\frac{(1\,\Delta\theta)(-\hat{e}_r)}{\Delta\theta} = -\hat{e}_r. \tag{16}$$

Then (14) becomes $\dot{\hat{e}}_\theta = -\dot{\theta}\hat{e}_r$, and (13) gives

$$\mathbf{a}(t) = \ddot{r}\hat{e}_r + \dot{r}\dot{\theta}\hat{e}_\theta + \dot{r}\dot{\theta}\hat{e}_\theta + r\ddot{\theta}\hat{e}_\theta - r\dot{\theta}^2\hat{e}_r$$

or,

$$\boxed{\mathbf{a}(t) = (\ddot{r}-r\dot{\theta}^2)\hat{e}_r + (r\ddot{\theta}+2\dot{r}\dot{\theta})\hat{e}_\theta.} \tag{17}$$

Of the four terms constituting $\mathbf{a}$, two with well-known names are the *centripetal acceleration* $-r\dot{\theta}^2\hat{e}_r$, and the *Coriolis acceleration* $2\dot{r}\dot{\theta}\hat{e}_\theta$.*

Although the formulas (12) and (17) are of great importance, for our purposes (especially in Chapter 16 on scalar and vector field theory) the most important part of this discussion concerns the differentiation of the base vectors, for in non-Cartesian coordinate systems the base vectors generally depend on one or more of the coordinates so whenever we differentiate a vector we need to face up to differentiating the base vectors. In particular, for plane polar coordinates we found that the base vectors $\hat{e}_r$ and $\hat{e}_\theta$ vary with θ, but not with r, and that

$$\boxed{\frac{d\hat{e}_r}{d\theta} = \hat{e}_\theta \quad\text{and}\quad \frac{d\hat{e}_\theta}{d\theta} = -\hat{e}_r.} \tag{18a,b}$$

The method that we used to derive equations (18a,b) will be called the **difference quotient method**. That method is appealing pedagogically because it relies on fundamentals -- the difference quotient definitions (10) and (15) together with the graphical representation of the various terms. However, it is heuristic rather than rigorous. For instance, expressing $\mathbf{cd} = (1\Delta\theta)(-\hat{e}_r)$, in (16), seemed to be correct but was not proved.

*The latter is named after the French engineer *G. Coriolis* (1792–1843), but it might be mentioned that the expressions $a_r = \ddot{r} - r\dot{\theta}^2$ and $a_\theta = r\ddot{\theta} + 2\dot{r}\dot{\theta}$ were derived by Euler in his book *Theoria Motus Corporum Solidorum seu Rigidorum*, published in 1765.

Thus, we now present a different line of approach which *is* rigorous, and which we will call the **transform method**. The idea is to express $\hat{e}_r$ and $\hat{e}_\theta$ (temporarily) in terms of the alternative Cartesian basis $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, the point being that $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors and hence readily differentiated. Specifically,

$$\hat{e}_r = \cos\theta\hat{\mathbf{i}} + \sin\theta\hat{\mathbf{j}}, \tag{19a}$$

$$\hat{e}_\theta = -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}. \tag{19b}$$

[The fact that $\hat{e}_r$ and $\hat{e}_\theta$ depend only on θ, as stated in (5), is seen to follow from the absence of any r dependence in the right-hand sides of (19).] Taking $d/d\theta$, we have

$$\frac{d\hat{e}_r}{d\theta} = -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}, \tag{20a}$$

$$\frac{d\hat{e}_\theta}{d\theta} = -\cos\theta\hat{\mathbf{i}} - \sin\theta\hat{\mathbf{j}}. \tag{20b}$$

Again, we emphasize that this step was simple because $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors. With the differentiation completed, we now return to the polar base vectors by the inverse relations

$$\hat{\mathbf{i}} = \cos\theta\hat{e}_r - \sin\theta\hat{e}_\theta, \tag{21a}$$

$$\hat{\mathbf{j}} = \sin\theta\hat{e}_r + \cos\theta\hat{e}_\theta. \tag{21b}$$

That is, putting (21) into (20),

$$\begin{aligned}\frac{d\hat{e}_r}{d\theta} &= -\sin\theta(\cos\theta\hat{e}_r - \sin\theta\hat{e}_\theta) + \cos\theta(\sin\theta\hat{e}_r + \cos\theta\hat{e}_\theta)\\ &= (\sin^2\theta + \cos^2\theta)\hat{e}_\theta = \hat{e}_\theta,\end{aligned}$$

and

$$\begin{aligned}\frac{d\hat{e}_\theta}{d\theta} &= -\cos\theta(\cos\theta\hat{e}_r - \sin\theta\hat{e}_\theta) - \sin\theta(\sin\theta\hat{e}_r + \cos\theta\hat{e}_\theta)\\ &= -(\cos^2\theta + \sin^2\theta)\hat{e}_r = -\hat{e}_r,\end{aligned}$$

as before.

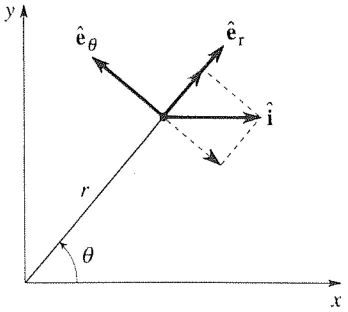


Figure 4. Expanding $\hat{\mathbf{i}}$.

In case the source of the relations (19) and (21) was not clear, let us elaborate. Consider (21a), for instance. Since $\{\hat{e}_r, \hat{e}_\theta\}$ is an orthonormal basis we can, according to the important equation (24) in Section 9.9, expand $\hat{\mathbf{i}}$ as $\hat{\mathbf{i}} = (\hat{\mathbf{i}}\cdot\hat{e}_r)\hat{e}_r + (\hat{\mathbf{i}}\cdot\hat{e}_\theta)\hat{e}_\theta = (1)(1)\cos\theta\hat{e}_r + (1)(1)\cos(\theta+\frac{\pi}{2})\hat{e}_\theta = \cos\theta\hat{e}_r - \sin\theta\hat{e}_\theta$ (see Fig. 4). Or, we could simply break $\hat{\mathbf{i}}$ into its $\hat{e}_r$ and $\hat{e}_\theta$ components graphically as indicated in Fig. 4. Similarly for (21b), (19a), and (19b).

We call the latter method the transform method because the "transform equations" (19) take us from $\hat{e}_r$, $\hat{e}_\theta$ to $\hat{\mathbf{i}}, \hat{\mathbf{j}}$. Then the $d/d\theta$ operation is readily carried out [in equations (20)] because $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors. With the results in hand, we then use the "inverse equations" (21) to return to $\hat{\mathbf{e}}_r$, $\hat{\mathbf{e}}_\theta$. If this idea is unclear, we suggest that you review Section 5.1.

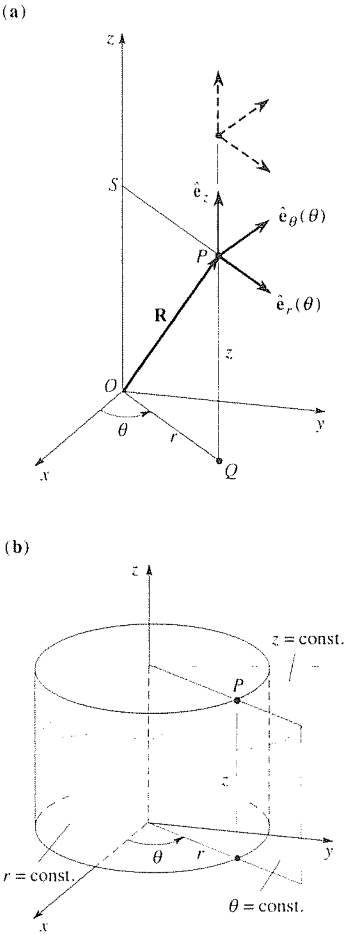


Figure 5. Cylindrical coordinates.

14.6.2. Cylindrical coordinates. The cylindrical coordinate system r, θ, z with orthonormal base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$, shown in Fig. 5, is essentially the plane polar coordinate system that we have just discussed, with the coordinate z and the corresponding base vector $\hat{\mathbf{e}}_z$ added. Observe that the z in r, θ, z is the same as the Cartesian z variable. Similarly, $\hat{\mathbf{e}}_z$ is identical to $\hat{\mathbf{k}}$, and we would write $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{k}}$, except that $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ looks better.

As z is varied, with r and θ fixed, the point P (in Fig. 5a) moves along a vertical line. The $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ triad translates (see the dashed arrows in Fig. 5a) but does not rotate. Thus, since those vectors do not vary either in magnitude or in direction, they do not vary with z. Similarly, as r is varied, with θ and z fixed, P slides along the line SP. The triad translates but does not rotate so $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ do not vary with r. Finally, $\hat{\mathbf{e}}_z$ does not vary with θ but $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ do, as discussed in Section 14.6.1 above. The upshot is that $\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta)$, $\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta)$, and $\hat{\mathbf{e}}_z = \hat{\mathbf{k}}$ is a constant vector, and the derivatives $d\hat{\mathbf{e}}_r/d\theta$ and $d\hat{\mathbf{e}}_\theta/d\theta$ are as given in (18).

As we did for plane polar coordinates, let us develop expressions for the position vector, velocity, and acceleration. It might be tempting to think that the position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r + \theta\hat{\mathbf{e}}_\theta + z\hat{\mathbf{e}}_z$, by analogy with $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$. No, an expression for $\mathbf{R}$ is obtained from Fig. 5a: $\mathbf{R} = \mathbf{OQ} + \mathbf{QP}$, where $\mathbf{OQ} = r\hat{\mathbf{e}}_r$ and $\mathbf{QP} = z\hat{\mathbf{e}}_z$. Thus,

$$\boxed{\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z,} \tag{22}$$

which is the same as (6), but with $z\hat{\mathbf{e}}_z$ added. Remember, whereas $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ are functions of θ, $\hat{\mathbf{e}}_z$ is a *constant vector*; at every point P, $\hat{\mathbf{e}}_z$ has the same magnitude and the same direction. Thus, if $r = r(t)$, $\theta = \theta(t)$, and $z = z(t)$, where t is the time, then the velocity $\mathbf{v} = \dot{\mathbf{R}}$ and the acceleration $\mathbf{a} = \ddot{\mathbf{R}}$ are given by the formulas

$$\boxed{\mathbf{v}(t) = \dot{\mathbf{R}}(t) = \dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta + \dot{z}\hat{\mathbf{e}}_z} \tag{23}$$

and

$$\boxed{\mathbf{a}(t) = \ddot{\mathbf{R}}(t) = (\ddot{r} - r\dot{\theta}^2)\hat{\mathbf{e}}_r + (r\ddot{\theta} + 2\dot{r}\dot{\theta})\hat{\mathbf{e}}_\theta + \ddot{z}\hat{\mathbf{e}}_z.} \tag{24}$$

That is, we merely append $\dot{z}\hat{\mathbf{e}}_z$ to (12) and $\ddot{z}\hat{\mathbf{e}}_z$ to (17).

EXAMPLE 1. Let a particle move according to

$$r = \text{constant}, \qquad \theta = \omega t, \qquad z = vt,$$

where ω and v are constants, and $0 \le t < \infty$. Evidently, the path of motion is helical. From (23) and (24),

$$\mathbf{v}(t) = r\omega\hat{\mathbf{e}}_\theta + v\hat{\mathbf{e}}_z, \tag{25a}$$

$$\mathbf{a}(t) = -r\omega^2\hat{\mathbf{e}}_r. \tag{25b}$$

Observe that $\mathbf{v}(t)$ is *not* constant, even though its components $r\omega$ and v are constant, because $\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta(t))$ varies with t. [Indeed, if $\mathbf{v}(t)$ *were* constant, then $\mathbf{a}(t) = \dot{\mathbf{v}}(t)$ would be zero, not $-r\omega^2\hat{\mathbf{e}}_r$.] ▮

Under what circumstances is one coordinate system to be favored over another? One important guideline is the shape of the region involved. For instance, cylindrical coordinates are most convenient for studying fluid flow in a cylindrical pipe, and Cartesian coordinates are most convenient for studying the electric field in a rectangular prism. That is, *a particular coordinate system is convenient if the region is bounded by constant-coordinate surfaces.* Thus, Cartesian coordinates are convenient if the region is bounded by constant-x, constant-y, and constant-z planes, that is, if it is a rectangular prism. How about cylindrical coordinates? Constant-r, -θ, and -z surfaces are shown in Fig. 5b, where the $\theta =$ constant surface is known as a *meridional plane*. Thus, some representative regions for which cylindrical coordinates are especially convenient are as follows (sketch them).

$$\begin{aligned}&\text{Soup can}: \quad 0 \le r \le r_1, \quad 0 \le \theta < 2\pi, \quad 0 \le z \le z_1.\\ &\text{Coaxial cable}: \quad r_1 \le r \le r_2, \quad 0 \le \theta < 2\pi, \quad 0 \le z \le z_1.\\ &\text{Wedge}: \quad 0 \le r \le r_1, \quad 0 \le \theta < \theta_1, \quad 0 \le z \le z_1.\end{aligned}$$

14.6.3. Spherical coordinates. Finally, we present the spherical coordinates ρ, ϕ, θ with their respective orthonormal unit base vectors $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ shown in Fig. 6. Some authors use the notation r, ϕ, θ, but we prefer to emphasize the difference between the spherical coordinate ρ and the cylindrical coordinate r by using different letters. What difference? Well, $\rho = \sqrt{x^2 + y^2 + z^2}$ is the distance from the origin to P, whereas $r = \sqrt{x^2 + y^2}$ is not (unless $z = 0$); it is the perpendicular distance from P to the z axis. The angle θ, however, *is* the same in the two coordinate systems, cylindrical and spherical.

Observe from Fig. 6b that the $\rho =$ constant surface through P is the *spherical* surface of radius ρ centered at the origin, the $\phi =$ constant surface is *conical*, and the $\theta =$ constant surface is a *meridional plane*.

In general, ρ, ϕ, and θ may be positive, negative, or zero, but we choose, in this book, always to have $\rho \ge 0$ so that ρ is the *distance* from the origin to P. In addition, we normally limit ϕ and θ so that $0 \le \phi \le \pi$ and $0 \le \theta < 2\pi$ since those ranges permit us to reach any point in 3-space in a unique manner.

From Fig. 6a we can see that ρ, ϕ, θ are related to the reference x, y, z coordinates according to

$$\begin{aligned}x &= \rho \sin\phi \cos\theta\\ y &= \rho \sin\phi \sin\theta \qquad\qquad (26a)\\ z &= \rho \cos\phi,\end{aligned}$$

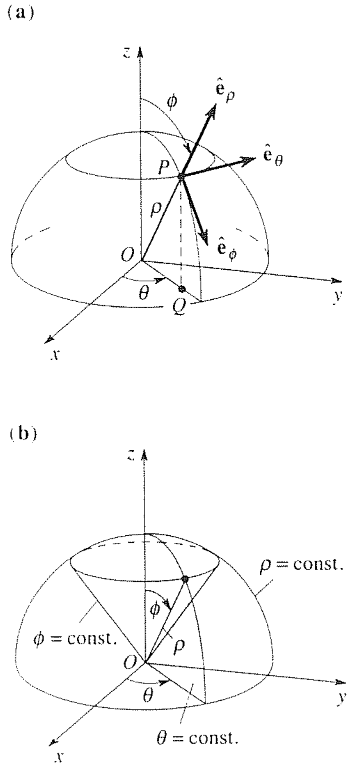


Figure 6. Spherical coordinates.

or, solving (26a) for ρ, ϕ, and θ,

$$\begin{aligned} \rho &= \sqrt{x^2+y^2+z^2}, \\ \phi &= \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}}, \\ \theta &= \tan^{-1}\frac{y}{x}, \end{aligned} \tag{26b}$$

where θ is undefined on the z axis (i.e., for $x = y = 0$) and ϕ is undefined at the origin. That is, on the positive z axis ϕ is zero and on the negative z axis ϕ is π so at the origin ϕ is discontinuous and undefined.

EXAMPLE 2. Given the point $(x, y, z) = (2, -1, 3)$, find ρ, ϕ, θ (such that $\rho \geq 0$, $0 \leq \phi \leq \pi$, and $0 \leq \theta < 2\pi$ as noted above). From (26b), $\rho = \sqrt{2^2+(-1)^2+3^2} = \sqrt{14}$, $\phi = \cos^{-1}(3/\sqrt{14}) = 36.70°$, and $\theta = \tan^{-1}(-1/2) = 153.4°$ *and* 333.4° (within the interval $0 \leq \theta < 360°$). To choose between the latter two values, put the computed ρ, ϕ, θ values back into (26a) and see if they give the original x, y, z values: $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 333.4°$ gives $(x, y, z) = (2, -1, 3)$ but $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 153.4°$ does not. Thus, $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 333.4°$. ▮

As emphasized earlier in this section, it is essential, in working with this coordinate system, to know the "space derivatives" of the base vectors, namely, $\partial/\partial\rho$, $\partial/\partial\phi$, $\partial/\partial\theta$ of each of the three base vectors. From Fig. 6a we see that if we hold ϕ and θ fixed and vary ρ, the $\hat{e}_\rho, \hat{e}_\phi, \hat{e}_\theta$, triad translates parallel to itself as P moves along the radial line. Thus, $\hat{e}_\rho, \hat{e}_\phi, \hat{e}_\theta$ do not vary with ρ so $\partial\hat{e}_\rho/\partial\rho = \partial\hat{e}_\phi/\partial\rho = \partial\hat{e}_\theta/\partial\rho = \mathbf{0}$. Next, imagine varying ϕ, with ρ and θ fixed. Then (Fig. 6a) $\hat{e}_\theta$ moves parallel to itself, and hence does not change, whereas $\hat{e}_\rho$ and $\hat{e}_\phi$ both rotate. Finally, if we vary θ, with ρ and ϕ fixed, we see that all three base vectors rotate. The upshot is that the dependence of the base vectors on the coordinates is as follows:

$$\hat{e}_\rho = \hat{e}_\rho(\phi, \theta), \qquad \hat{e}_\phi = \hat{e}_\phi(\phi, \theta), \qquad \hat{e}_\theta = \hat{e}_\theta(\theta). \tag{27}$$

Thus, there are five nonzero derivatives to work out. To do so, we could use either of the two methods presented above in Section 14.6.1, the difference quotient method or the transform method. However, the calculations are harder than for the case of plane polar coordinates so we merely state the results:

$$\boxed{\begin{array}{lll} \dfrac{\partial\hat{e}_\rho}{\partial\rho} = 0, & \dfrac{\partial\hat{e}_\rho}{\partial\phi} = \hat{e}_\phi, & \dfrac{\partial\hat{e}_\rho}{\partial\theta} = \sin\phi\,\hat{e}_\theta, \\[2ex] \dfrac{\partial\hat{e}_\phi}{\partial\rho} = 0, & \dfrac{\partial\hat{e}_\phi}{\partial\phi} = -\hat{e}_\rho, & \dfrac{\partial\hat{e}_\phi}{\partial\theta} = \cos\phi\,\hat{e}_\theta, \\[2ex] \dfrac{\partial\hat{e}_\theta}{\partial\rho} = 0, & \dfrac{\partial\hat{e}_\theta}{\partial\phi} = 0, & \dfrac{\partial\hat{e}_\theta}{\partial\theta} = -\sin\phi\,\hat{e}_\rho - \cos\phi\,\hat{e}_\phi, \end{array}} \tag{28}$$

and defer their derivation to Section 14.6.4, where we use a different, and simpler, approach.

Let us illustrate the use of (28) by deriving spherical-coordinate expressions for the velocity and acceleration, expressions analogous to those obtained for plane polar and cylindrical coordinates. We begin with the position vector $\mathbf{R} = \mathbf{OP} = \|\mathbf{OP}\|\,\hat{\mathbf{e}}_\rho$, that is,

$$\boxed{\mathbf{R} = \rho\hat{\mathbf{e}}_\rho.} \tag{29}$$

Since $\hat{\mathbf{e}}_\rho = \hat{\mathbf{e}}_\rho(\phi(t), \theta(t))$, differentiation with respect to t gives

$$\begin{aligned}\mathbf{v} = \dot{\mathbf{R}} &= \dot\rho\,\hat{\mathbf{e}}_\rho + \rho\left(\frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi}\frac{d\phi}{dt} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta}\frac{d\theta}{dt}\right)\\ &= \dot\rho\,\hat{\mathbf{e}}_\rho + \rho(\dot\phi\,\hat{\mathbf{e}}_\phi + \dot\theta\sin\phi\,\hat{\mathbf{e}}_\theta)\end{aligned}$$

or

$$\boxed{\mathbf{v}(t) = \dot\rho\,\hat{\mathbf{e}}_\rho + \rho\dot\phi\,\hat{\mathbf{e}}_\phi + \rho\dot\theta\sin\phi\,\hat{\mathbf{e}}_\theta.} \tag{30}$$

One more differentiataion (Exercise 7) gives

$$\boxed{\begin{aligned}\mathbf{a}(t) = {}&(\ddot\rho - \rho\dot\phi^2 - \rho\dot\theta^2\sin^2\phi)\hat{\mathbf{e}}_\rho + (\rho\ddot\phi + 2\dot\rho\dot\phi - \rho\dot\theta^2\sin\phi\cos\phi)\hat{\mathbf{e}}_\phi\\ &+ (\rho\ddot\theta\sin\phi + 2\dot\rho\dot\phi\sin\phi + 2\rho\dot\theta\dot\phi\cos\phi)\hat{\mathbf{e}}_\theta\end{aligned}} \tag{31}$$

for the acceleration.

14.6.4. Omega method. Here, we will develop an interesting alternative method for deriving the space derivatives of base vectors, and use that method to derive equations (28).

Consider a rigid (i.e., nondeformable) body $\mathcal{B}$ undergoing an arbitrary motion through 3-space, and let $\mathbf{A}$ be any fixed vector within $\mathcal{B}$ (Fig. 7). That is, $\mathbf{A}$ is a vector from one material point in $\mathcal{B}$ to another so $\|\mathbf{A}\|$ is constant with time because $\mathcal{B}$ is rigid. That is not to say that $\mathbf{A}$ is a constant vector, for a vector is comprised of a magnitude and a direction, and the direction of $\mathbf{A}$ varies with the time t, in general, as $\mathcal{B}$ tumbles through space. Thus, $\mathbf{A} = \mathbf{A}(t)$. It might help to think of $\mathcal{B}$ as a potato, say, and $\mathbf{A}$ as a needle pushed into the potato.

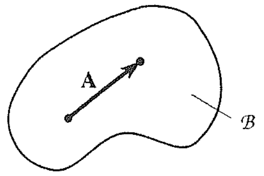


Figure 7. Fixed vector in $\mathcal{B}$.

Since $\|\mathbf{A}\|$ is constant ,

$$\|\mathbf{A}\|^2 = \mathbf{A}\cdot\mathbf{A} = \text{constant} \tag{32}$$

as well. Differentiating (32) with respect to t gives [recall (7c) in Section 14.5] $\dot{\mathbf{A}}\cdot\mathbf{A} + \mathbf{A}\cdot\dot{\mathbf{A}} = 2\mathbf{A}\cdot\dot{\mathbf{A}} = 0$ so

$$\mathbf{A}\cdot\dot{\mathbf{A}} = 0. \tag{33}$$

Equation (33) makes sense because if $\dot{\mathbf{A}}$ were *not* orthogonal to $\mathbf{A}$, then it would have a nonzero component *along* $\mathbf{A}$, which component would correspond to a nonzero rate of stretching or shrinking of $\mathbf{A}$, in violation of the assumption that $\mathbf{A}$ is of fixed length.

It follows from (33) (see Exercise 12) that there exists a vector $\boldsymbol{\Omega}_1$ such that

$$\dot{\mathbf{A}} = \boldsymbol{\Omega}_1 \times \mathbf{A}. \tag{34}$$

Similarly, if $\mathbf{B}$ is any other fixed vector in $\mathcal{B}$ then there exists a vector $\boldsymbol{\Omega}_2$ such that

$$\dot{\mathbf{B}} = \boldsymbol{\Omega}_2 \times \mathbf{B}. \tag{35}$$

Since $\mathcal{B}$ is rigid it follows that

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\alpha = \text{constant} \tag{36}$$

because $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are constant, as is the angle α between them. Differentiating (36) with respect to t gives

$$\dot{\mathbf{A}} \cdot \mathbf{B} + \mathbf{A} \cdot \dot{\mathbf{B}} = 0 \tag{37}$$

or, using (34) and (35),

$$\begin{aligned} (\boldsymbol{\Omega}_1 \times \mathbf{A}) \cdot \mathbf{B} + \mathbf{A} \cdot (\boldsymbol{\Omega}_2 \times \mathbf{B}) &= 0, \\ \boldsymbol{\Omega}_1 \times \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \times \boldsymbol{\Omega}_2 \cdot \mathbf{B} &= 0, \qquad \text{[by (7) in Section 14.4]} \end{aligned}$$

so

$$(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2) \times \mathbf{A} \cdot \mathbf{B} = 0. \tag{38}$$

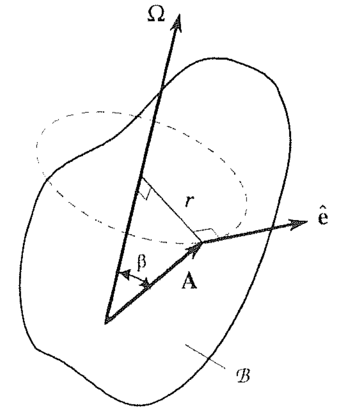


Figure 8. The angular velocity $\boldsymbol{\Omega}$.

Since $\mathbf{B}$ is arbitrary, it follows from (38) (Exercise 13) that $(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2) \times \mathbf{A} = \mathbf{0}$. Finally, since $\mathbf{A}$ is arbitrary it follows from the latter (Exercise 14) that $\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2 = \mathbf{0}$. Thus, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$, and we can say that for any fixed vector $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ in $\mathcal{B}$ we have $\dot{\mathbf{A}} = \boldsymbol{\Omega} \times \mathbf{A}$, $\dot{\mathbf{B}} = \boldsymbol{\Omega} \times \mathbf{B}$, $\dot{\mathbf{C}} = \boldsymbol{\Omega} \times \mathbf{C}$, and so on. Or, stated more concisely,

$$\boxed{\dot{\mathbf{A}} = \boldsymbol{\Omega} \times \mathbf{A},} \tag{39}$$

where $\mathbf{A}$ is *any* fixed vector in $\mathcal{B}$.

To understand what $\boldsymbol{\Omega}$ is, let us write out (39) as

$$\dot{\mathbf{A}} = (\|\boldsymbol{\Omega}\|)\,(\|\mathbf{A}\| \sin\beta)\,\hat{\mathbf{e}} = \Omega\, r\, \hat{\mathbf{e}}, \tag{40}$$

where β, r, and $\hat{\mathbf{e}}$ are shown in Fig. 8; $\hat{\mathbf{e}}$ is tangent to the circle of radius r (shown as dashed). But (40) is the familiar result, encountered in physics, that the velocity of a point moving in a circle of radius r with angular velocity Ω is Ωr in the tangential direction. That is, Ω is the instantaneous angular velocity of $\mathcal{B}$. Using the right-hand rule to give it a direction, Ω is therefore the **angular velocity vector** of $\mathcal{B}$; $\boldsymbol{\Omega}$ need not be constant with time but Fig. 8 and equation (39) hold at each instant.

Before using (39) to derive (28), let us consider the simpler case of cylindrical coordinates. Imagine the base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ (Fig. 5a) as welded steel rods within a body $\mathcal{B}$ (which could be built up around the rods using modeling clay), and imagine that body to be in motion according to $r = r(t)$, $\theta = \theta(t)$, and $z = z(t)$, where $r(t)$, $\theta(t)$, and $z(t)$ are arbitrary. Then what is the angular velocity $\boldsymbol{\Omega}$ of the body? If we vary r only (holding θ and z fixed), then the body translates, with no angular velocity. Similarly if we vary z (holding r and θ fixed). But if we vary θ, then $\mathcal{B}$ undergoes an angular velocity $\dot{\theta}(t)$ or, using the right-hand rule to make it into a vector, $\dot{\theta}(t)\hat{\mathbf{e}}_z$. Since angular velocity is a vector, we can get the total angular velocity by adding these three contributions: $\boldsymbol{\Omega} = \mathbf{0} + \mathbf{0} + \dot{\theta}\,\hat{\mathbf{e}}_z$ so

$$\boldsymbol{\Omega} = \dot{\theta}\,\hat{\mathbf{e}}_z. \tag{41}$$

To determine the space derivatives of $\hat{\mathbf{e}}_r$, say, let $\mathbf{A}$ be $\hat{\mathbf{e}}_r$ (which does have fixed length) in (39). Then (39) gives

$$\frac{d\hat{\mathbf{e}}_r}{dt} = \boldsymbol{\Omega} \times \hat{\mathbf{e}}_r = \dot{\theta}\,\hat{\mathbf{e}}_z \times \hat{\mathbf{e}}_r = \dot{\theta}\,\hat{\mathbf{e}}_\theta. \tag{42}$$

On the other hand, we can use chain differentiation to write

$$\begin{aligned} \frac{d}{dt}\,\hat{\mathbf{e}}_r(r(t), \theta(t), z(t)) &= \frac{\partial \hat{\mathbf{e}}_r}{\partial r}\frac{dr}{dt} + \frac{\partial \hat{\mathbf{e}}_r}{\partial \theta}\frac{d\theta}{dt} + \frac{\partial \hat{\mathbf{e}}_r}{\partial z}\frac{dz}{dt} \\ &= \dot{r}\frac{\partial \hat{\mathbf{e}}_r}{\partial r} + \dot{\theta}\frac{\partial \hat{\mathbf{e}}_r}{\partial \theta} + \dot{z}\frac{\partial \hat{\mathbf{e}}_r}{\partial z}. \end{aligned} \tag{43}$$

Next, we equate the right-hand sides of (42) and (43) and use the fact that $r(t), \theta(t), z(t)$ (and hence $\dot{r}, \dot{\theta}, \dot{z}$) are arbitrary:

$$\dot{r}\mathbf{0} + \dot{\theta}\hat{\mathbf{e}}_\theta + \dot{z}\mathbf{0} = \dot{r}\frac{\partial \hat{\mathbf{e}}_r}{\partial r} + \dot{\theta}\frac{\partial \hat{\mathbf{e}}_r}{\partial \theta} + \dot{z}\frac{\partial \hat{\mathbf{e}}_r}{\partial z} \tag{44}$$

so

$$\frac{\partial \hat{\mathbf{e}}_r}{\partial r} = \mathbf{0}, \qquad \frac{\partial \hat{\mathbf{e}}_r}{\partial \theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{\partial \hat{\mathbf{e}}_r}{\partial z} = \mathbf{0}. \tag{45}$$

That is, since $\dot{r}(t), \dot{\theta}(t), \dot{z}(t)$ are arbitrary we can set $\dot{r} = 1, \dot{\theta} = 0, \dot{z} = 0$ in (44), and learn that $\partial\hat{\mathbf{e}}_r/\partial r = \mathbf{0}$. Then we can set $\dot{r} = 0, \dot{\theta} = 1, \dot{z} = 0$, and so on. [Of course we didn't really need to include the $\partial\hat{\mathbf{e}}_r/\partial r$ and $\partial\hat{\mathbf{e}}_r/\partial z$ terms in (43) because we knew, from our discussion in Section 14.6.1 and 14.6.2 that $\hat{\mathbf{e}}_r$ is a function of θ alone, but we included those terms to show that we don't *need* to know that information in advance.]

Similarly, if we use $\mathbf{A} = \hat{\mathbf{e}}_\theta$ we obtain $\partial\hat{\mathbf{e}}_\theta/\partial r = \mathbf{0}$, $\partial\hat{\mathbf{e}}_\theta/\partial\theta = -\hat{\mathbf{e}}_r$, and $\partial\hat{\mathbf{e}}_\theta/\partial z = \mathbf{0}$, and if we use $\mathbf{A} = \hat{\mathbf{e}}_z$ we obtain $\partial\hat{\mathbf{e}}_z/\partial r = \partial\hat{\mathbf{e}}_z/\partial\theta = \partial\hat{\mathbf{e}}_z/\partial z = \mathbf{0}$.

Understand that we *imagine* the representative point P as undergoing a motion $r(t)$, $\theta(t)$, $z(t)$ simply in order to be able to use this method, which we call the **omega method** because of the prominent role played by $\boldsymbol{\Omega}$.

Turning to spherical coordinates, what is $\boldsymbol{\Omega}$ in this case? Referring to Fig. 6a, we see that varying ρ (holding ϕ and θ fixed) results in no angular velocity of the base vector triad, only translation; varying ϕ (holding ρ and θ fixed) results in an angular velocity $\dot{\phi}(t)\hat{\mathbf{e}}_\theta$; and varying θ (holding ρ and ϕ fixed) results in an angular velocity about the z axis, $\dot{\theta}(t)\hat{\mathbf{k}}$. Thus,

$$\begin{aligned}\boldsymbol{\Omega} &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\hat{\mathbf{k}} \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left[(\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\rho)\hat{\mathbf{e}}_\rho + (\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\phi)\hat{\mathbf{e}}_\phi + (\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta\right] \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left[\cos\phi\hat{\mathbf{e}}_\rho + \cos\left(\frac{\pi}{2}+\phi\right)\hat{\mathbf{e}}_\phi + 0\hat{\mathbf{e}}_\theta\right] \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left(\cos\phi\hat{\mathbf{e}}_\rho - \sin\phi\hat{\mathbf{e}}_\phi\right). \end{aligned} \tag{46}$$

Let us determine the space derivatives of $\hat{\mathbf{e}}_\rho$, say. Let $\mathbf{A} = \hat{\mathbf{e}}_\rho$ in (39). Then, on the one hand

$$\begin{aligned}\frac{d\hat{\mathbf{e}}_\rho}{dt} = \boldsymbol{\Omega}\times\hat{\mathbf{e}}_\rho &= \left[\dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}(\cos\phi\hat{\mathbf{e}}_\rho - \sin\phi\hat{\mathbf{e}}_\phi)\right]\times\hat{\mathbf{e}}_\rho \\ &= \dot{\phi}\hat{\mathbf{e}}_\phi + \dot{\theta}\sin\phi\hat{\mathbf{e}}_\theta, \end{aligned} \tag{47}$$

and on the other hand

$$\frac{d}{dt}\hat{\mathbf{e}}_\rho(\rho(t),\phi(t),\theta(t)) = \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho}\dot{\rho} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi}\dot{\phi} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta}\dot{\theta}, \tag{48}$$

and comparison of (47) and (48) gives

$$\frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho} = 0, \qquad \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi} = \hat{\mathbf{e}}_\phi, \qquad \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta} = \sin\phi\hat{\mathbf{e}}_\theta,$$

in agreement with (28). Then, letting $\mathbf{A}$ be $\hat{\mathbf{e}}_\phi$ and $\hat{\mathbf{e}}_\theta$, in turn, we can obtain the remaining space derivatives, which steps we leave for the exercises.

Observe that $\boldsymbol{\Omega}$ is determined by the coordinate system: for plane polar and cylindrical coordinates $\boldsymbol{\Omega} = \dot{\theta}\hat{\mathbf{k}}$, and for spherical coordinates $\boldsymbol{\Omega} = \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\hat{\mathbf{k}}$. There is no need to use the omega method for Cartesian coordinates since $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ are constant vectors so $\partial\hat{\mathbf{i}}/\partial x = \partial\hat{\mathbf{i}}/\partial y = \partial\hat{\mathbf{i}}/\partial z = \partial\hat{\mathbf{j}}/\partial x = \cdots = \mathbf{0}$, but it is worth mentioning that for that coordinate system $\boldsymbol{\Omega} = \mathbf{0}$.

Closure. In this section we derive expressions for the velocity and acceleration vectors in plane polar, cylindrical, and spherical coordinates, but that discussion merely serve as a common application thread and is not, in itself, the focal point. The focal point is the plane polar, cylindrical, and spherical coordinate systems: the definition of the coordinates by means of a sketch (e.g., Fig. 5), the relationship between those coordinates and the reference Cartesian coordinates (e.g., $x = r\cos\theta$, $y = r\sin\theta$, and $z = z$, the Cartesian coordinate z being identical to the cylindrical coordinate z), the expression for the position vector (e.g., $\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$ in cylindrical coordinates), and the space derivatives of the base vectors [e.g., equations (28) for spherical coordinates]. In fact, most of our attention is devoted to the derivation of the derivatives of the base vectors, and three distinct methods are presented: the difference quotient method, the transform method, and the omega method. These derivatives will show up extensively in Chapter 16, but you don't need to derive them each time they arise; that has already been done here, and you can use the results, listed in (18) for cylindrical coordinates and (28) for spherical coordinates.

EXERCISES 14.6

1. Determine r, θ, z corresponding to each given point $P = (x, y, z)$. Give θ both in radians ($0 \le \theta < 2\pi$) and in degrees ($0° \le \theta < 360°$).

(a) $(2, 2, 3)$
(b) $(2, -1, 0)$
(c) $(-3, 2, 1)$
(d) $(0, 0, 5)$
(e) $(6, -1, 3)$
(f) $(-1, -2, 6)$
(g) $(-1, 5, 1)$
(h) $(6, 0, 4)$

2. Determine ρ, ϕ, θ corresponding to each given point $P = (x, y, z)$ such that $\rho \ge 0$, $0 \le \phi \le \pi$, $0 \le \theta < 2\pi$. Give ϕ and θ both in radians and in degrees.

(a) $(2, 0, 0)$
(b) $(0, 3, 0)$
(c) $(0, 5, 0)$
(d) $(1, 2, 3)$
(e) $(6, -3, 1)$
(f) $(0, -5, 0)$
(g) $(2, 3, 5)$
(h) $(2, 3, -5)$
(i) $(-2, 3, 5)$
(j) $(2, -3, 5)$
(k) $(-1, -2, 1)$
(l) $(4, -2, -3)$

3. (a) Derive the expressions

$$\begin{aligned} \hat{e}_\rho &= \sin\phi(\cos\theta\hat{i} + \sin\theta\hat{j}) + \cos\phi\hat{k}, \\ \hat{e}_\phi &= \cos\phi(\cos\theta\hat{i} + \sin\theta\hat{j}) - \sin\phi\hat{k}, \\ \hat{e}_\theta &= -\sin\theta\hat{i} + \cos\theta\hat{j}. \end{aligned} \tag{3.1}$$

(b) Derive the "reverse" expressions

$$\begin{aligned} \hat{i} &= \sin\phi\cos\theta\hat{e}_\rho + \cos\phi\cos\theta\hat{e}_\phi - \sin\theta\hat{e}_\theta, \\ \hat{j} &= \sin\phi\sin\theta\hat{e}_\rho + \cos\phi\sin\theta\hat{e}_\phi + \cos\theta\hat{e}_\theta, \\ \hat{k} &= \cos\phi\hat{e}_\rho - \sin\phi\hat{e}_\phi, \end{aligned} \tag{3.2}$$

either directly or by solving (3.1) for $\hat{i}, \hat{j}, \hat{k}$. If you choose the latter, you may use computer software.

4. Compute $d\mathbf{A}/dt$ in each case.

(a) $\mathbf{A} = \sin\theta\hat{e}_r - r\theta^2\hat{e}_\theta + r\hat{e}_z$; $r = t^2$, $\theta = 3t$
(b) $\mathbf{A} = t\hat{e}_r$; $r = 4$, $\theta = 2t$
(c) $\mathbf{A} = r\hat{e}_r$; $r = 6 + \sin t$, $\theta = \cos t$
(d) $\mathbf{A} = \rho^2\hat{e}_\rho$; $\rho = t$, $\phi = 2t$, $\theta = 3t$
(e) $\mathbf{A} = \hat{e}_\phi$; $\rho = 1 + t$, $\phi = t^2$, $\theta = \sin 2t$
(f) $\mathbf{A} = \hat{e}_\rho + t\hat{e}_\theta$; $\rho = 1$, $\phi = t$, $\theta = t^2$.

5. A particle moves in 3-space according to the given functions $x(t), y(t), z(t)$. Determine its velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}(t)$ in terms of $\hat{e}_r, \hat{e}_\theta, \hat{e}_z$, and t, for $t \ge 0$.

(a) $x = t^2$, $y = 2$, $z = 0$
(b) $x = 2t$, $y = 5t$, $z = 3t$
(c) $x = t^2 - t$, $y = t^2$, $z = -3t$
(d) $x = 3\cos t$, $y = 3\sin t$, $z = 6t$
(e) $x = 1$, $y = 4t$, $z = 2$
(f) $x = \sin t$, $y = \cos t$, $z = \sin t$

6. (a)–(f) Same as Exercise 5, but find $\mathbf{v}$ and $\mathbf{a}$ in terms of $\hat{e}_\rho, \hat{e}_\phi, \hat{e}_\theta$, and t, for $t \ge 0$. You may use any formulas in Exercise 3.

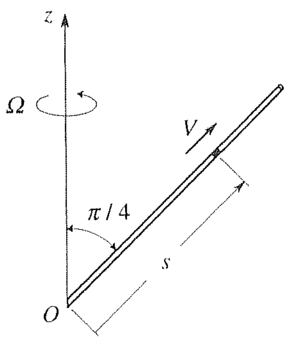
7. Evaluate each of the following by the difference quotient method, and show that your result agrees with that given in (28).

(a) $\partial\hat{e}_\rho/\partial\phi$
(b) $\partial\hat{e}_\rho/\partial\theta$
(c) $\partial\hat{e}_\phi/\partial\phi$
(d) $\partial\hat{e}_\phi/\partial\theta$
(e) $\partial\hat{e}_\theta/\partial\theta$

8. Derive (31) from (30), with the help of (28).

9. Beginning at O at time $t = 0$, a bead moves at a constant speed V through a straight hollow tube, which rotates at constant angular velocity Ω about the z axis so as to sweep out a cone of half angle $\pi/4$, as shown. Compute the velocity $\mathbf{v}$, and the acceleration $\mathbf{a}$ of the particle, in terms of s, V, and Ω, using

(a) Cartesian coordinates
(b) cylindrical coordinates
(c) spherical coordinates



NOTE: Let us orient the x, y, z coordinate frame so that the tube lies in the x, z plane at $t = 0$.

10. Beginning at the "north pole" ($\phi = 0$), a particle spirals down the surface of a sphere of radius ρ such that the spherical coordinates ϕ and θ vary as $\phi = \Omega_1 t$, $\theta = \Omega_2 t$. Find the magnitude of the velocity and acceleration vectors at the instant when the particle crosses the "equator."

11. Observe, in (28), that none of the derivatives of $\hat{\mathbf{e}}_\rho$ contain an $\hat{\mathbf{e}}_\rho$ component, that none of the derivatives of $\hat{\mathbf{e}}_\phi$ contain an $\hat{\mathbf{e}}_\phi$ component, and that none of the derivatives of $\hat{\mathbf{e}}_\theta$ contain an $\hat{\mathbf{e}}_\theta$ component. The same pattern holds in (18). Explain why this pattern is more than coincidence.

12. Given vectors $\mathbf{u}$ and $\mathbf{v}$ in 3-space, such that $\mathbf{u} \cdot \mathbf{v} = 0$, where $\|\mathbf{u}\| \neq 0$, show that there exists a vector $\mathbf{\Omega}$ such that $\mathbf{v} = \mathbf{\Omega} \times \mathbf{u}$. HINT: Write $\mathbf{v} = \mathbf{\Omega} \times \mathbf{u}$ as three scalar equations on the components $\Omega_1, \Omega_2, \Omega_3$ of $\mathbf{\Omega}$, and solve.

13. Show that if $\mathbf{u} \cdot \mathbf{v} = 0$ for all vectors $\mathbf{v}$, then it must be true that $\mathbf{u} = \mathbf{0}$.

14. Show that if $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ for all vectors $\mathbf{v}$, then it must be true that $\mathbf{u} = \mathbf{0}$.

15. Use the omega method to evaluate the three space derivatives of the spherical coordinate base vector

(a) $\hat{\mathbf{e}}_\phi$ (b) $\hat{\mathbf{e}}_\theta$

Chapter 14 Review

In this review let us limit attention to results that we urge you to know without needing to look them up.

Section 14.2

$$\|\mathbf{u} \times \mathbf{v}\| = \text{area of } \mathbf{u}, \mathbf{v} \text{ parallelogram}$$

Section 14.3 If $\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}$ and $\mathbf{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}$, then

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

Section 14.4

$$|\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}| = \text{volume of } \mathbf{u}, \mathbf{v}, \mathbf{w} \text{ parallelepiped}$$

$$\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

$$\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \mathbf{u} \times \mathbf{v} \cdot \mathbf{w}$$

Section 14.6

Polar coordinates: Know Fig. 1.

Coordinates r, θ and base vectors $\hat{\mathbf{e}}_r(\theta), \hat{\mathbf{e}}_\theta(\theta)$.

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \end{aligned}$$

Position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r$.

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

Cylindrical coordinates: Know Fig. 5a.

Coordinates r, θ, z and base vectors $\hat{\mathbf{e}}_r(\theta), \hat{\mathbf{e}}_\theta(\theta), \hat{\mathbf{e}}_z$.

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta \\ z &= z \end{aligned}$$

Position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$.

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

Spherical coordinates: Know Fig. 6a.

Coordinates ρ, ϕ, θ and base vectors $\hat{\mathbf{e}}_\rho(\phi, \theta), \hat{\mathbf{e}}_\phi(\phi, \theta), \hat{\mathbf{e}}_\theta(\theta)$.

$$\begin{aligned} x &= \rho\sin\phi\cos\theta \\ y &= \rho\sin\phi\sin\theta \\ z &= \rho\cos\phi \end{aligned}$$

Position vector is $\mathbf{R} = \rho\hat{\mathbf{e}}_\rho$.

Space derivatives of the base vectors given by (28); no need to memorize these.

Chapter 15

Curves, Surfaces, and Volumes

15.1 Introduction

In applications such as the calculation of moments of inertia, particle dynamics, and the gravitational force field induced by a given arrangement of mass, we need to know how to represent curves, surfaces, and volumes in 3-space, and how to obtain, from those representations, expressions for the differential arc length ds along a curve, the area element dA on a surface, and the volume element dV. We begin with space curves and integrals along those curves.

15.2 Curves and Line Integrals

Curves in space are, of course, of great importance. In mechanics, for example, the relationship between the forces acting on a point mass and the trajectory of the mass (i.e., the space curve developed by the motion) is of critical interest.

We met curves, indirectly, in Section 14.5, when we considered the variation of a vector function $\mathbf{u}(\tau)$ with τ. For, in general, if the tail of the vector $\mathbf{u}(\tau)$ is fixed, then the head of $\mathbf{u}(\tau)$ generates a space curve as the parameter τ is varied. Here, we shift our focus from the vector function $\mathbf{u}(\tau)$ to the space curve itself.

15.2.1. Curves. Let a Cartesian x, y, z coordinate system be specified, and let

$$x = x(\tau), \qquad y = y(\tau), \qquad z = z(\tau) \tag{1}$$

be continuous functions of a real parameter τ over a closed interval $[a, b]$, that is, over $a \leq \tau \leq b$. The points $P(\tau) \equiv (x(\tau), y(\tau), z(\tau))$ for $a \leq \tau \leq b$ are said to constitute a **curve** joining the endpoints $P(a)$ and $P(b)$, and we call (1) a **parametrization** of the curve.* If the endpoints coincide the curve is called a

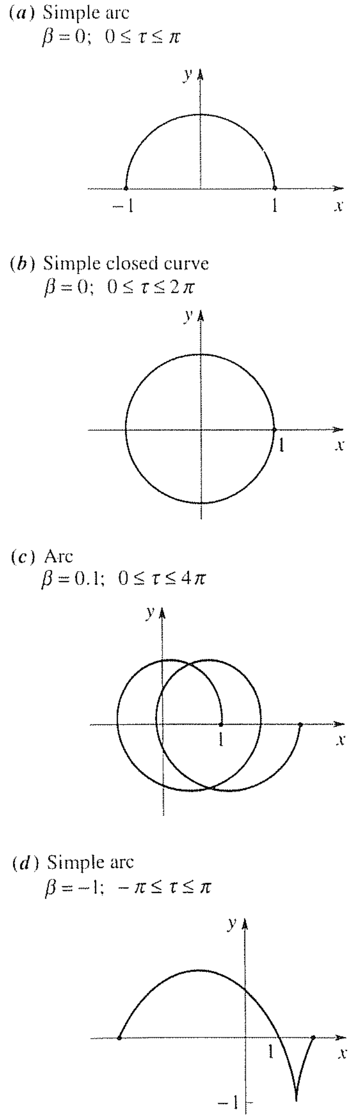
*If each of $x(\tau)$, $y(\tau)$, and $z(\tau)$ is constant, the point set $P(\tau)$ is a single point and is called a *point curve*. Although we limit ourselves, in this section, to curves in 3-space, the definition given here may be extended to curves in n-space.

closed curve; if not, it is called an **arc**. If the tail of the position vector

$$\mathbf{R} = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}} \tag{2}$$

is fixed at the origin, then the head of $\mathbf{R}(\tau)$ generates the curve as τ varies from a to b.

As indicated in Section 14.5 (see the discussion associated with Fig. 3 therein), the derivative $\mathbf{R}'(\tau)$ (if it exists and is nonzero) is a **tangent vector** to the curve at the point $P(\tau)$.


(*a*) Simple arc
$\beta = 0;\ 0 \le \tau \le \pi$

(*b*) Simple closed curve
$\beta = 0;\ 0 \le \tau \le 2\pi$

(*c*) Arc
$\beta = 0.1;\ 0 \le \tau \le 4\pi$

(*d*) Simple arc
$\beta = -1;\ -\pi \le \tau \le \pi$

Figure 1. Several cases of the curve (3).

EXAMPLE 1. Suppose that

$$x = \beta\tau + \cos\tau, \qquad y = \sin\tau, \qquad z = 0, \tag{3}$$

where β is a constant. Then $\mathbf{R}(\tau) = (\beta\tau + \cos\tau)\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}}$. Since $z = 0$, the curve lies in a plane (the x, y plane in this case) and is therefore called a *plane curve*. Choosing different values of β and different τ intervals, we obtain different curves. Four such cases are illustrated in Fig. 1. ▮

We classify a curve C as **simple** if it does not intersect itself. If C is an arc it will be a *simple arc* if $\tau_1 \neq \tau_2$ implies that $P(\tau_1) \neq P(\tau_2)$ whenever τ_1 and τ_2 are in $[a, b]$; if C is a closed curve it will be a **simple closed curve** if $\tau_1 \neq \tau_2$ implies that $P(\tau_1) \neq P(\tau_2)$ whenever τ_1 and τ_2 are in $[a, b)$.* For instance, the curves in Fig. 1a, b, d are simple, and the one in Fig. 1c is not.

Further, we say that C is **smooth** if it possesses a tangent vector that varies continuously along the length of C, and it is **piecewise smooth** if it is comprised of finitely many smooth segments, end to end. For instance, the curves in Fig. 1a, b, c are smooth, and the one in Fig. 1d is piecewise smooth.

In Example 1 we were given the parametric equations of the curve from which we were able to produce its graph. Sometimes we know the curve, but are not given parametric equations for it. In the next two examples we illustrate the determination of parametric equations for a given curve.

EXAMPLE 2. Find parametric equations for the curve C that is the intersection of the two planes

$$\begin{aligned} x - y + 2z &= 4, \\ 2x + y - z &= 2. \end{aligned} \tag{4}$$

C is the solution set of equations (4) so apply Gauss elimination and find that $z = \tau$ (arbitrary), $y = -2 + \frac{5}{3}\tau$, and $x = 2 - \frac{1}{3}\tau$. Indeed, these are the desired parametric equations of C, where $-\infty < \tau < \infty$. Or, in 3-tuple vector form, $(x, y, z) =$

*$[a, b)$ means $a \le \tau < b$. We write $[a, b)$ rather than $[a, b]$ because the endpoints $P(a)$ and $P(b)$ coincide: $P(a) = P(b)$.

$(2 - \frac{1}{3}\tau, -2 + \frac{5}{3}\tau, \tau) = (2, -2, 0) + \tau(-\frac{1}{3}, \frac{5}{3}, 1)$, where $(2, -2, 0)$ is from the origin to a particular point *on* C and $(-\frac{1}{3}, \frac{5}{3}, 1)$ is a vector *along* C. ▮

EXAMPLE 3. Find parametric equations for the curve C that is the ellipse

$$\left(\frac{x}{2}\right)^2 + \left(\frac{y}{3}\right)^2 = 1 \tag{5}$$

in the x, y plane. If we had a *circle* $x^2 + y^2 = 1$ we could use $x = \cos\tau$ and $y = \sin\tau$, that is, $x = r\cos\theta$ and $y = r\sin\theta$, where $r = 1$ and where we use θ as our parameter τ. Adapting that idea to (5), let $\frac{x}{2} = \cos\tau$ and $\frac{y}{3} = \sin\tau$. Thus, parametric equations for the ellipse (5) are

$$x = 2\cos\tau, \qquad y = 3\sin\tau, \tag{6}$$

where $0 \le \tau < 2\pi$, say. [Whether or not we add $z = 0$ to (6) depends on whether we are considering C as a curve in the x, y plane or in x, y, z space.] ▮

Observe that the parametrization of a given curve is not unique. In Example 3, for instance, the alternative parametrizations $x = 2\sin\tau$, $y = 3\cos\tau$ and $x = 2\cos\tau^2$, $y = 3\sin\tau^2$ work just as well.

One case worth emphasizing is the parametrization of a straight line from a given point $P_1(x_1, y_1, z_1)$ to another given point $P_2(x_2, y_2, z_2)$. We claim that

$$\boxed{\begin{aligned} x &= x_1 + (x_2 - x_1)\tau, \\ y &= y_1 + (y_2 - y_1)\tau, \\ z &= z_1 + (z_2 - z_1)\tau \end{aligned}} \qquad (0 \le \tau \le 1) \tag{7}$$

always works. Obviously, the right-hand sides of (7) give (x_1, y_1, z_1) when $\tau = 0$, and (x_2, y_2, z_2) when $\tau = 1$. Not quite as obvious is the fact that the curve thus parametrized is *straight*. To verify that it is straight, write the position vector from the origin to any point on the curve as

$$\mathbf{R}(\tau) = [x_1 + (x_2 - x_1)\tau]\hat{\mathbf{i}} + [y_1 + (y_2 - y_1)\tau]\hat{\mathbf{j}} + [z_1 + (z_2 - z_1)\tau]\hat{\mathbf{k}}.$$

Since the tangent vector to the curve,

$$\mathbf{R}'(\tau) = (x_2 - x_1)\hat{\mathbf{i}} + (y_2 - y_1)\hat{\mathbf{j}} + (z_2 - z_1)\hat{\mathbf{k}},$$

is a constant vector, it follows that the curve parametrized by (7) is, indeed, a straight line.

15.2.2. Arc length. Next, we consider the arc length $s(\tau)$ of a curve C, from some point $P(\tau_0)$ on C to any other point $P(\tau)$ on C. We know what we mean by the length of a straight line, but what is the length of a curved arc? As is done so often in mathematics, we use the limit concept to define this "new" quantity as the limit of a sequence of "old" quantities.* Specifically, we approximate the curve by a system of linear segments, as in Fig. 2, and sum the lengths of those segments. Then we repeat the process using a finer system of linear segments, and so on. If the endpoints of any single linear segment (such as AB in Fig. 2) are located with respect to the origin of some coordinate system by position vectors $\mathbf{R}_A$ and $\mathbf{R}_B$, then the length of the segment is

$$\|\mathbf{R}_B - \mathbf{R}_A\| = \|\Delta\mathbf{R}\| = \sqrt{\Delta\mathbf{R}\cdot\Delta\mathbf{R}} = \sqrt{\frac{\Delta\mathbf{R}}{\Delta\tau}\cdot\frac{\Delta\mathbf{R}}{\Delta\tau}}\,\Delta\tau$$

so in differential form we have

$$\boxed{ds = \sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\,d\tau.} \tag{8}$$

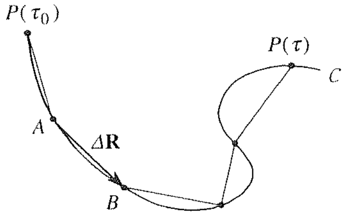


Figure 2. Approximation of C by linear segments.

Thus, the limiting process of summation described above evidently leads to the Riemann integral

$$\boxed{s(\tau) = \int_{\tau_0}^{\tau} \sqrt{\mathbf{R}'(t)\cdot\mathbf{R}'(t)}\,dt,} \tag{9}$$

which we adopt as our *definition* of the **arc length** of C, from $P(\tau_0)$ to $P(\tau)$, provided of course that the integral exists (i.e., is convergent). It can be shown that the arc length of a given arc is independent of the location and orientation of the reference Cartesian coordinate system and is also independent of the particular choice of parametrization $x(\tau), y(\tau), z(\tau)$. CAUTION: The integral in (9) gives the arc length only if $\tau \geq \tau_0$; if $\tau < \tau_0$, then the integral is negative and gives the negative of the arc length between $P(\tau_0)$ and $P(\tau)$.

EXAMPLE 4. As a simple illustration, use (9) to compute the arc length of the (semicircular) curve shown in Fig. 1a. Then

$$\begin{aligned}\mathbf{R}(\tau) &= x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} = \cos\tau\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}},\\ \mathbf{R}'(\tau) &= -\sin\tau\hat{\mathbf{i}} + \cos\tau\hat{\mathbf{j}},\\ \mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau) &= \sin^2\tau + \cos^2\tau = 1\end{aligned}$$

so (9) is simply

$$s = \int_0^{\pi} \sqrt{1}\,dt = \pi. \quad ▮$$

*Recall, for instance, that the derivative is defined (in the calculus) as the limit of a difference quotient, the integral is defined as the limit of a sequence of Riemann sums, an infinite series is ordinarily defined as the limit of a sequence of partial sums, and so on. It is interesting that the limit is the fundamental and unifying concept of the calculus, yet it was not clarified until the nineteenth century, through the work of Cauchy (1789–1857), Weierstrass (1815–1897), and others, long after the birth of the calculus.

15.2.3. Line integrals. Consider a given curve C and a function of $f(x, y, z)$ defined along C (perhaps off of C as well). Measuring arc length s from one endpoint, say A (Fig. 3), divide C into n arcs by specifying points P_0 ($= A$), $P_1, P_2, \dots, P_{n-1}, P_n$ ($= B$) along C. Let the division be chosen arbitrarily, provided that the P_j points are spaced and numbered so that the arc length from A to P_j is less than arc length from A to P_k if $j < k$. Denote the arcs as $C_1, C_2, \dots, C_n$, where the endpoints of C_j are P_{j-1} and P_j. On each arc C_j choose, arbitrarily, some point Q_j that is anywhere between the endpoints of C_j, or *at* one of the endpoints, and form the sum

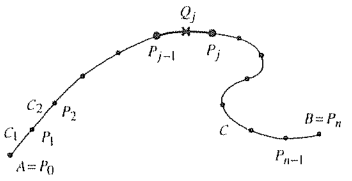


Figure 3. Partition of C.

$$J_n = \sum_{j=1}^{n} f(Q_j)\Delta s_j, \tag{10}$$

where Δs_j is the arc length of C_j. The choice of the P_j's and Q_j's defines a **partition** of C, and we call the largest Δs_j the **norm** of the partition. Actually, we introduce not just one partition but a sequence of them such that the norm of the nth partition tends to zero as $n \to \infty$. If the corresponding sequence of sums $J_1, J_2, \dots$ converges to a limit, we call that limit the **line integral** of f over C, and write it as

$$\int_C f\,ds. \tag{11}$$

In this chapter we study integrals over curves, surfaces, and volumes so the terms curve integral, surface integral, and volume integral would appear to be natural choices. However, (11) is usually called a line integral, rather than a curve integral, so we will use that terminology.

The limit described above, and hence the integral (11), will indeed exist (i.e., converge) if f is continuous (or even piecewise continuous) along C and if C is piecewise smooth and simple.

From this definition of the line integral it can be shown that the following two important properties follow:

$$\int_C (\alpha f + \beta g) ds = \alpha \int_C f\,ds + \beta \int_C g\,ds \qquad \text{(linearity)} \tag{12a}$$

and

$$\int_C f\,ds = \int_{C_1} f\,ds + \int_{C_2} f\,ds, \tag{12b}$$

where α, β are any scalars, f, g are any two functions (continuous along C), and where "$C = C_1 + C_2$," that is, where C is divided into two parts, C_1 and C_2 (just as C was divided into n parts in Fig. 3).

In practice, however, we do not evaluate line integrals by seeking the limit of the sequence $J_1, J_2, \dots$. Rather, we introduce a coordinate system and proceed as follows. If C is parametrized with respect to some parameter τ by $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$, for $a \le \tau \le b$, then

$$\boxed{\int_C f(x, y, z)\,ds = \int_a^b f(x(\tau), y(\tau), z(\tau))\,\sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\,d\tau.} \tag{13}$$

EXAMPLE 5. *Mass of a Helical Wire.* Determine the mass

$$M = \int_C \sigma \, ds \tag{14}$$

of a wire that is in the shape of a curve C and that has a mass density σ (mass per unit length) that varies along C. In particular, suppose that C is comprised of N turns of a circular helix of radius a, defined parametrically by

$$x(\tau) = a\cos\tau, \quad y(\tau) = a\sin\tau, \quad z(\tau) = b\tau, \qquad (0 \le \tau \le 2N\pi) \tag{15}$$

where τ is actually the cylindrical-coordinate angle θ. Observe that in one turn of the helix (i.e., as τ increases by 2π) z increases by $2\pi b$, which quantity is known as the *pitch* of the helix. As b is diminished the helix becomes more and more compressed, collapsing to a circle of radius a in the x, y plane as $b \to 0$; conversely, it becomes more and more stretched out as b is increased. And let the density vary linearly with τ as $\sigma(\tau) = c\tau$ for some positive constant c. Then

$$\begin{aligned} \mathbf{R}(\tau) &= a\cos\tau\hat{\mathbf{i}} + a\sin\tau\hat{\mathbf{j}} + b\tau\hat{\mathbf{k}}, \\ \mathbf{R}'(\tau) &= -a\sin\tau\hat{\mathbf{i}} + a\cos\tau\hat{\mathbf{j}} + b\hat{\mathbf{k}} \end{aligned}$$

so (13) gives

$$M = \int_0^{2N\pi} c\tau\sqrt{a^2+b^2}\, d\tau = c\sqrt{a^2+b^2}\,\frac{(2N\pi)^2}{2}. \quad ▮$$

Closure. The key to space curves and curve integrals is in their parametric representation. We have considered space curves C defined parametrically by

$$x = x(\tau), \quad y = y(\tau), \quad z = z(\tau). \qquad (a \le \tau \le b)$$

The parameter τ might have a geometrical significance (as in Example 5, for instance, where τ amounted to the polar angle θ of an r, θ, z cylindrical coordinate system), it might represent time, or it might have no particular significance. Likewise, the differential arc length along C can be expressed in terms of τ as

$$ds = \sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\, d\tau,$$

where $\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}$ is a position vector from the origin of the reference Cartesian coordinate system to any point $P(\tau)$ on C. Finally, an integral $\int_C f\, ds$ along a curve C can, by parametrization, be reduced to an "ordinary integral" of a function of τ along a segment of a τ axis, as indicated in (13).

Computer software. Using *Maple*, you can obtain the graph of a curve C defined parametrically by

$$\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}} \qquad (a \le \tau \le b)$$

using the **spacecurve** command. For instance, to plot the helix

$$\mathbf{R}(\tau) = 3\cos\tau\hat{\mathbf{i}} + 3\sin\tau\hat{\mathbf{j}} + 20\tau\hat{\mathbf{k}} \qquad (0 \le \tau \le 25)$$

enter

```
with(plots):
```

and return, to access the spacecurve command. Then enter

```
spacecurve([3 * cos (t), 3 * cos (t), 20 * t], t = 0..25);
```

and return. Two problems: first is that the graph is kinky because the default number of points used is only 50 and that is not enough for approximately four (25 divided by 2π) turns of the helix and, second, axes are not shown. To increase the number of points to 500, say, use the option numpoints = 500, and to show the labeled coordinate axes use the option axes = FRAMED or axes = BOXED. Thus, use

```
spacecurve([3 * cos (t), 3 * cos (t), 20 * t], t = 0..25,
numpoints = 500, axes = BOXED);
```

for instance.

EXERCISES 15.2

1. Determine a parametrization for each of the following curves: the intersection of

(a) the planes $x - 2y + z = 4,\ \ 2x + y - z = 0$
(b) the planes $4x - 3y + 5z = 1,\ \ x + y - 2z = 6$
(c) the planes $2x + y + z = 3,\ \ x + 4z = 5$
(d) the planes $x - y + z = 0,\ \ 5x - y + z = 2$
(e) the planes $2x + y - z = 1,\ \ 2y + z = 4$
(f) the plane $x + y + 2z = 5$ and the circular cylinder $x^2 + y^2 = 4$
(g) the plane $4x - y + z = 3$ and the circular cylinder $x^2 + z^2 = 1$
(h) the plane $x - y + 2z = 1$ and the elliptic cylinder $x^2 + 4y^2 = 4$
(i) surfaces $2x - y - 3z = 5$ and $z = x^2 + y + 1$, between $(0, -2, -1)$ and $(2, -4, 1)$
(j) surfaces $x - y^2 + z = 0$ and $x - y + 2z = 0$, between $(6, 2, -2)$ and $(1, 1, 0)$
(k) surfaces $x = y^2 + z^2$ and $z = xy$, between $(0, 0, 0)$ and $(1, -1/\sqrt{2}, -1/\sqrt{2})$

2. Give a parametrization of the straight line connecting
(a) $(5, -1, 2)$ and $(2, 0, 6)$ (b) $(1, 2, 3)$ and $(3, 7, -4)$
(c) $(8, 1, 7)$ and $(0, 28, 12)$ (d) $(1, 0, 0)$ and $(-6, 0, 0)$

3. Let each of the following $\mathbf{R}(\tau)$'s, for $0 \le \tau < \infty$, define a space curve. In each case, find the arc length s as a function of the parameter τ, namely, $s(\tau)$ if $\tau = 0$ is taken to correspond to $s = 0$. NOTE: Naturally, $\cos\tau\hat{\mathbf{i}}$, for instance, means $(\cos\tau)\hat{\mathbf{i}}$.

(a) $\hat{\mathbf{i}} + \tau\hat{\mathbf{j}} + \tau^2\hat{\mathbf{k}}$ (b) $\cos\tau\hat{\mathbf{i}} + 4\tau\hat{\mathbf{j}} + \sin\tau\hat{\mathbf{k}}$
(c) $\tau\hat{\mathbf{i}} - 3\tau\hat{\mathbf{j}} + 5(\tau + 4)\hat{\mathbf{k}})$ (d) $\cos\tau(\hat{\mathbf{i}} + \hat{\mathbf{j}}) - \sqrt{2}\sin\tau\hat{\mathbf{k}}$
(e) $\hat{\mathbf{i}} - 7\tau^2\hat{\mathbf{j}} + 3\hat{\mathbf{k}}$ (f) $\sin\tau\hat{\mathbf{i}} + \cos\tau\hat{\mathbf{j}} - \cos\tau\hat{\mathbf{k}}$

4. (*Serret–Frenet formulas*) The arc length s along a space curve C is itself a convenient parameter for the parametrization of C. Letting τ be s, we have

$$\mathbf{R}(s) = x(s)\hat{\mathbf{i}} + y(s)\hat{\mathbf{j}} + z(s)\hat{\mathbf{k}}. \tag{4.1}$$

In what follows there are problems and statements. We ask you to follow along with the statements and to respond only to problems introduced by the italicized word *show*.

(a) Since $d\mathbf{R}$ is tangent to C, $d\mathbf{R}/ds$ must be a tangent vector to C. *Show* why the **tangent vector**

$$\frac{d\mathbf{R}}{ds} \equiv \hat{\mathbf{T}} \tag{4.2}$$

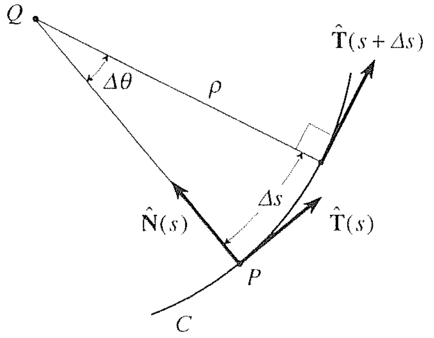
is a unit vector.
(b) Since $\hat{\mathbf{T}}(s)$ is a unit vector, $\hat{\mathbf{T}} \cdot \hat{\mathbf{T}} = 1$. Differentiating that equation with respect to s, *show* that $\hat{\mathbf{T}} \cdot \hat{\mathbf{T}}' = 0$. Hence, $\hat{\mathbf{T}}'$ is either perpendicular to $\hat{\mathbf{T}}'$ or else it is $\mathbf{0}$ so we can express

$$\frac{d\hat{\mathbf{T}}}{ds} = \kappa\hat{\mathbf{N}} \qquad (\kappa \geq 0) \tag{4.3}$$

where $\hat{\mathbf{N}}$ is a unit normal vector perpendicular to $\hat{\mathbf{T}}$ and κ is a scalar multiplier, which can be considered as nonnegative without loss of generality. $\hat{\mathbf{N}}(s)$ is called the **principal normal** and $\kappa(s)$ is called the **curvature** of C. If the curve happens to be straight (in some neighborhood of the point in question), then $\hat{\mathbf{T}}(s)$ is a constant vector and $\hat{\mathbf{T}}'(s) = \mathbf{0}$ so that, from (4.3), $\kappa = 0$, as is reasonable since a straight line has no curvature. In that case $\hat{\mathbf{N}} = (1/\kappa)\hat{\mathbf{T}}'$ is undefined. The plane containing $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ is called the **osculating plane**.
(c) In (b), we stated that κ is the curvature, which means the numerical inverse of the **radius of curvature** ρ,

$$\kappa = \frac{1}{\rho}. \tag{4.4}$$

Show that it follows from (4.3) that κ is indeed the curvature. HINT: Use the accompanying figure (where Q is the local center of curvature corresponding to the point P on C) and a difference quotient approximation of $\hat{\mathbf{T}}'(s)$. Note from (4.4) that if the radius of curvature ρ is large then the curvature κ is small, and if ρ is small then κ is large. Finally, we introduce a third unit vector $\hat{\mathbf{B}}(s)$, the **binormal**, according to



$$\hat{\mathbf{B}} = \hat{\mathbf{T}} \times \hat{\mathbf{N}} \tag{4.5}$$

so that at each point P on C $\left\{\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}\right\}$ is a right-handed orthonormal set. The spatial orientation of the $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}$ triad varies, in general, as the representative point P [at the tip of $\mathbf{R}(s)$] moves along the curve C. Thus, in general, $\hat{\mathbf{T}}(s), \hat{\mathbf{N}}(s), \hat{\mathbf{B}}(s)$ will all be functions of s. Indeed, the variation of $\hat{\mathbf{T}}$ with s has already been given in (4.3). Let us obtain analogous expressions for $\hat{\mathbf{N}}'(s)$ and $\hat{\mathbf{B}}'(s)$.
(d) First $\hat{\mathbf{B}}'(s)$: From (4.5), *show* that $\hat{\mathbf{B}}' = \hat{\mathbf{T}} \times \hat{\mathbf{N}}'$. It follows from the latter that $\hat{\mathbf{B}}'$ is perpendicular to $\hat{\mathbf{T}}$ (also to $\hat{\mathbf{N}}'$, but that fact will not be used here). But $\hat{\mathbf{B}}$ is a unit vector so, by the same logic as given in (b) for $\hat{\mathbf{T}}$, $\hat{\mathbf{B}}'$ must be perpendicular to $\hat{\mathbf{B}}$ or be $\mathbf{0}$. Thus, $\hat{\mathbf{B}}'$ is perpendicular to the $\hat{\mathbf{B}}$, $\hat{\mathbf{T}}$ plane, or else it is $\mathbf{0}$ so it must be expressible in the form

$$\frac{d\hat{\mathbf{B}}}{ds} = \nu\hat{\mathbf{N}}, \qquad (\nu \geq 0) \tag{4.6}$$

where the scalar multiple ν in (4.6) is called the **torsion** of C at P because it is a measure of the rate at which the curve twists out of the osculating plane; if the curve is straight we define $\nu = 0$. NOTE: The torsion is generally denoted as τ, not ν, in the literature, but we have already used τ extensively as a parameter and prefer to use a different letter for the torsion.
(e) Finally, differentiating $\hat{\mathbf{N}} = \hat{\mathbf{B}} \times \hat{\mathbf{T}}$ (for recall that $\left\{\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}\right\}$ is a right-handed orthonormal set) *show* that

$$\frac{d\hat{\mathbf{N}}}{ds} = -\kappa\hat{\mathbf{T}} - \nu\hat{\mathbf{B}}. \tag{4.7}$$

Equations (4.3), (4.6), and (4.7), namely,

$$\boxed{\begin{aligned} \frac{d\hat{\mathbf{T}}}{ds} &= \kappa\hat{\mathbf{N}}, \\ \frac{d\hat{\mathbf{N}}}{ds} &= -\kappa\hat{\mathbf{T}} - \nu\hat{\mathbf{B}}, \\ \frac{d\hat{\mathbf{B}}}{ds} &= \nu\hat{\mathbf{N}}, \end{aligned}} \tag{4.8a,b,c}$$

are known as the **Serret–Frenet formulas**.* The latter are fundamental to the study of space curves for the following reason. If $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}$ are expressed in terms of their scalar components then (4.8) is a system of nine linear first-order differential

*These formulas were obtained by *Frederic-Jean Frenet* (1816–1900) in his dissertation (1847) and, independently, by *Joseph Alfred Serret* (1819-1885) in 1851.

equations. If $\kappa(s)$ and $\nu(s)$ are prescribed continuous functions, then that system admits a solution for $\hat{\mathbf{T}}(s), \hat{\mathbf{N}}(s), \hat{\mathbf{B}}(s)$ that is unique to within the specification of $\hat{\mathbf{T}}(0), \hat{\mathbf{N}}(0), \hat{\mathbf{B}}(0)$. With $\hat{\mathbf{T}}(s)$ thus determined, we can determine $\mathbf{R}(s)$ and hence the space curve $\mathcal{C}$, by integrating (4.2). For details, we refer the interested reader to J. J. Stoker, *Differential Geometry* (New York: Wiley–Interscience, 1969) or D. J. Struik, *Differential Geometry*, 2nd ed. (Reading, MA: Addison–Wesley, 1961).

5. First, read Exercise 4. Given the circular helix

$$\mathbf{R}(\tau) = a\cos\tau\hat{\mathbf{i}} + a\sin\tau\hat{\mathbf{j}} + b\tau\hat{\mathbf{k}} \qquad (a, b > 0,\ 0 \le \tau < \infty) \tag{5.1}$$

derive the results

$$\begin{aligned} s(\tau) &= \sqrt{a^2+b^2}\,\tau, \\ \hat{\mathbf{T}}(\tau) &= c(-a\sin\tau\hat{\mathbf{i}} + a\cos\tau\hat{\mathbf{j}} + b\hat{\mathbf{k}}), \\ \hat{\mathbf{N}}(\tau) &= -\cos\tau\hat{\mathbf{i}} - \sin\tau\hat{\mathbf{j}}, \\ \hat{\mathbf{B}}(\tau) &= c(b\sin\tau\hat{\mathbf{i}} - b\cos\tau\hat{\mathbf{j}} + a\hat{\mathbf{k}}), \\ \kappa(\tau) &= a/(a^2+b^2), \\ \rho(\tau) &= (a^2+b^2)/a, \\ \nu(\tau) &= -b/(a^2+b^2), \end{aligned} \tag{5.2}$$

where $c \equiv 1/\sqrt{a^2+b^2}$. HINT: Don't forget to use chain differentiation. For example, $\hat{\mathbf{T}} = \dfrac{d\mathbf{R}}{ds} = \dfrac{d\mathbf{R}}{d\tau}\dfrac{d\tau}{ds}$.

6. In Exercise 5 we see that for a circular helix the curvature $\kappa(\tau)$ and torsion $\nu(\tau)$ are constant along the helix. In this exercise let us work in the opposite direction. Specifically, let κ and ν be constants in the Frenet–Serret equations (4.8) and show, by deriving it, that if κ and ν are not both zero then the solution of (4.8) is

$$\begin{aligned} \hat{\mathbf{N}} &= \mathbf{C}_1 \sin\omega s + \mathbf{C}_2\cos\omega s, \\ \hat{\mathbf{T}} &= \frac{\kappa}{\omega}\mathbf{C}_2\sin\omega s - \frac{\kappa}{\omega}\mathbf{C}_1\cos\omega s + \nu\mathbf{C}_3, \\ \hat{\mathbf{B}} &= \frac{\nu}{\omega}\mathbf{C}_2\sin\omega s - \frac{\nu}{\omega}\mathbf{C}_1\cos\omega s - \kappa\mathbf{C}_3, \end{aligned} \tag{6.1}$$

where $\omega \equiv \sqrt{\kappa^2+\nu^2}$, and $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ are the (vector) constants of integration. Then show that integration of (4.2) gives

$$\mathbf{R} = \mathbf{C}_4\cos\omega s + \mathbf{C}_5\sin\omega s + \mathbf{C}_6 s + \mathbf{C}_7. \tag{6.2}$$

Finally, verify that (5.1), in Exercise 5, is indeed of the form (6.2).

7. (*Velocity and acceleration of a particle*) First, read Exercise 4. Let the velocity of a moving particle be given by a position vector $\mathbf{R} = \mathbf{R}(t)$, where t is the time. We wish to obtain expressions for the velocity $\mathbf{v} = d\mathbf{R}/dt$ and the acceleration $\mathbf{a} = d\mathbf{v}/dt = d^2\mathbf{R}/dt^2$ in terms of $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}, \kappa$, and ν.

(a) With the help of chain differentiation (namely, $\dfrac{d(\,)}{dt} = \dfrac{d(\,)}{ds}\dfrac{ds}{dt} = v\dfrac{d(\,)}{ds}$, where $\dfrac{ds}{dt} = v = \|\mathbf{v}\|$ is the *speed*), derive the formulas

$$\boxed{\begin{aligned} \mathbf{v} &= v\hat{\mathbf{T}}, \\ \mathbf{a} &= \frac{dv}{dt}\hat{\mathbf{T}} + \kappa v^2\hat{\mathbf{N}}. \end{aligned}} \tag{7.1a,b}$$

NOTE: Equation (7.1b) is simple and informative. Since $\mathbf{a}$ is a linear combination of $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ it lies in the osculating plane. The first term on the right is simply the linear acceleration, the acceleration that would occur if the particle moved in a straight line with variable speed $v(t)$. The second term is directed toward the instantaneous center of curvature (the point Q in the figure in Exercise 4) and its magnitude is κv^2, or v^2/ρ. This is the familiar *centripetal acceleration* that we meet in physics when studying circular motion. Note further that we can now appreciate that the Serret–Frenet formulas (4.8) play the same role as the formulas for the spatial derivatives of the non-Cartesian base vectors (such as $\partial\hat{\mathbf{e}}_\rho/\partial\phi$, $\partial\hat{\mathbf{e}}_\phi/\partial\theta$, and so on) in Section 14.6. For in deriving (7.1b) from (7.1a) we needed to evaluate $d\hat{\mathbf{T}}/dt$. Chain differentiation gave $(d\hat{\mathbf{T}}/ds)(ds/dt)$, and then the Serret–Frenet formula (4.8a) gave $d\hat{\mathbf{T}}/ds$ as a linear combination of the base vectors $\hat{\mathbf{T}}, \hat{\mathbf{N}}$ and $\hat{\mathbf{B}}$.

(b) Actually, the *third* derivative $d^3\mathbf{R}/dt^3$ is also of interest in engineering applications (e.g., in cam design and in the analysis of mechanisms) and, quaintly, is called the **jerk**, $\mathrm{j}(t)$. Show, from (7.1b), that

$$\mathrm{j} = \left(\frac{\mathrm{d}^2\mathrm{v}}{\mathrm{dt}^2} - \kappa^2\mathrm{v}^3\right)\hat{\mathrm{T}} + 3\kappa\mathrm{v}\frac{\mathrm{dv}}{\mathrm{dt}}\hat{\mathrm{N}} - \kappa\nu\mathrm{v}^3\hat{\mathrm{B}}. \tag{7.2}$$

8. Compute the normal $\hat{\mathbf{N}}(x)$ for the plane curve $y = x^3$. Sketch the curve over $-2 \le x \le 2$, and show $\hat{\mathbf{N}}$ at a number of points, say, $x = \pm 0.01$ and $x = \pm 1.0$.

9. Let a curve in the x, y plane be defined by $y = y(x)$. Show that

$$\boxed{\kappa = \frac{1}{\rho} = \frac{|y''|}{(1+y'^2)^{3/2}},} \tag{9.1}$$

provided, of course, that $y(x)$ is twice differentiable.

10. Prove that $P(\tau) = (x(\tau), y(\tau), z(\tau))$ is a plane curve if and only if

$$\begin{vmatrix} x' & y' & z' \\ x'' & y'' & z'' \\ x''' & y''' & z''' \end{vmatrix} = 0, \tag{10.1}$$

where primes denote $d/d\tau$. HINT: Recall the general equation of a plane, $ax + by + cz = d$, where a, b, c, d are constants and a, b, c are not all zero. If $P(\tau)$ lies in a plane, then $ax(\tau) + by(\tau) + c(\tau) = d$.

11. Let each of the following $\mathbf{R}(\tau)$'s, for $0 \leq \tau < \infty$, define a space curve. Use the results in Exercise 10 to show whether or not the curve is a plane curve.

(a) $\tau\hat{\mathbf{i}} - 2\tau\hat{\mathbf{j}} + \tau^3\hat{\mathbf{k}}$
(b) $\tau^2\hat{\mathbf{i}} + \tau^2\hat{\mathbf{j}} - 3\tau\hat{\mathbf{k}}$
(c) $2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 5\tau\hat{\mathbf{k}}$
(d) $3\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}} + \cos\tau\hat{\mathbf{k}}$

12. (a)–(f) Obtain a computer plot of the curve defined in the corresponding part of Exercise 3, over $0 \leq \tau \leq 5$.

15.3 Double and Triple Integrals

Having studied curve integrals, we can turn to surface and volume integrals. Since these are multiple integrals, let us begin with a brief review of double and triple integrals.

15.3.1. Double integrals. Consider the double integral

$$I = \iint_{\mathcal{R}} f(x, y)\, dA, \tag{1}$$

where $\mathcal{R}$ is a region in the x, y plane. We assume that $\mathcal{R}$ is a closed, bounded region in the x, y plane,* that its boundary C is a closed piecewise smooth curve, and that $f(x, y)$ is defined on $\mathcal{R}$.

Let us lay down a rectangular grid on $\mathcal{R}$ consisting of a finite number of lines parallel to the coordinate axes, for example, as shown in Fig. 1. The N rectangles lying entirely within $\mathcal{R}$ (the shaded ones in Fig. 1) constitute a **partition** of $\mathcal{R}$, and the greatest of the dimensions $\Delta x_j, \Delta y_j$ for $j = 1, \ldots, N$ will be called the **norm** of the partition and will be denoted by the symbol $|p|$. Let (x_j, y_j) be an arbitrarily selected point in the jth partition rectangle for each $j = 1, 2, \ldots, N$. Then, denoting the area $\Delta x_j \Delta y_j$ as ΔA_j,

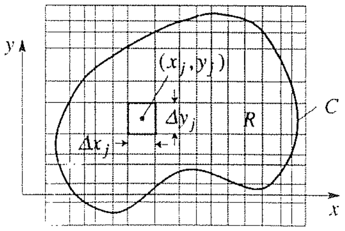


Figure 1. Partition of $\mathcal{R}$.

$$\sum_{j=1}^{N} f(x_j, y_j)\Delta A_j \tag{2}$$

is called the **Riemann sum** corresponding to the chosen partition and (x_j, y_j) points. The idea, in principle at least, is to compute the Riemann sum (2), then

*The term *closed region* is defined in Section 13.2.2. $\mathcal{R}$ being *bounded* means that it can be enclosed within a sufficiently large circle (or sphere in the three-dimensional case).

to introduce a finer partition (i.e., one with a smaller norm) and compute the new Riemann sum, and so on, such that the norm of the partitions tends to zero. Two things happen: the partitioning rectangles become arbitrarily small, and the unshaded area within $\mathcal{R}$ (Fig. 1) evidently tends to zero.

If the sequence of values of the Riemann sum thus generated converges to a unique limit, independent of the choice of the partition sequence and (x_j, y_j) points, that limit is, by definition, the **double integral** (1). That is,

$$\iint_{\mathcal{R}} f(x,y)\,dA \equiv \lim_{|p|\to 0} \sum f(x_j, y_j)\Delta A_j. \tag{3}$$

It can be shown that the limit will indeed exist if f is continuous in the closed region $\mathcal{R}$, that condition being sufficient, not necessary. If the limit does exist, we say that the integral *converges*, or *exists*, and that f is *integrable* on $\mathcal{R}$, Riemann-integrable to be more specific since the Riemann definition (3) is not the only definition of the double integral.

Double integrals admit the following properties, each of which corresponds to an analogous property for line integrals:

$$\iint_{\mathcal{R}} [\alpha f(x,y) + \beta g(x,y)]dA = \alpha \iint_{\mathcal{R}} f(x,y)\,dA + \beta \iint_{\mathcal{R}} g(x,y)\,dA, \tag{4a}$$

$$\iint_{\mathcal{R}} f(x,y)\,dA = \iint_{\mathcal{R}_1} f(x,y)\,dA + \iint_{\mathcal{R}_2} f(x,y)\,dA, \tag{4b}$$

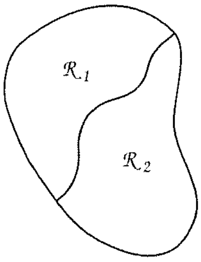


Figure 2. Subdivision of $\mathcal{R}$.

where α is a constant and all integrals are assumed to exist. The property (4a) is known as linearity. In (4b) $\mathcal{R}$ is divided (by a piecewise smooth curve) into two subregions, $\mathcal{R}_1$ and $\mathcal{R}_2$, as illustrated in Fig. 2. In addition to (4a) and (4b), we call attention to the **mean value theorem** for double integrals, which states that if $f(x,y)$ is continuous throughout the closed region $\mathcal{R}$, then there exists at least one point (x_0, y_0) in $\mathcal{R}$ such that

$$\boxed{\iint_{\mathcal{R}} f(x,y)\,dA = f(x_0, y_0)A,} \tag{4c}$$

where

$$A = \iint_{\mathcal{R}} dA \tag{5}$$

is the **area** of $\mathcal{R}$.

Besides serving to define the double integral on the left, (3) is useful from a computational standpoint since the Riemann sum (2), for a given partition, will provide an accurate approximation to the integral if the partition is sufficiently fine. [In practice, however, one normally uses approximations which are more efficient and more sophisticated than the simplest approximation provided by (2).]

Of course, we resort to numerical integration only when necessary, only when we are unable to carry out the evaluation analytically. How do we evaluate a double integral analytically? By dealing with it as an **iterated integral**, that is, by integrating on one variable at a time. The procedure is similar to that employed in differentiation. For example, in computing $u_{xy} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x}u\right)$, first we differentiate u with respect to x, holding y fixed, and then we differentiate the resulting function $\partial u/\partial x$ with respect to y. Analogously, if we recall that $\Delta A_j = \Delta x_j \Delta y_j$ in (2), this fact suggests that we write $\iint_{\mathcal{R}} f(x,y)\,dA$ as $\iint_{\mathcal{R}} f(x,y)\,dx\,dy$, and that we contemplate two *successive* integrations, first on x and then on y. Thus, if $\mathcal{R}$ is the region shown in Fig. 3a, let us tentatively express

$$\iint\limits_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2}\left\{\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx\right\}dy$$

or, omitting the parentheses for brevitiy,

$$\iint\limits_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2}\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx\,dy. \tag{6}$$

In words, we integrate across the (bold) horizontal sliver in Fig. 3a, from $x_1(y)$ on the left to $x_2(y)$ on the right, and then we sweep the sliver from bottom ($y = y_1$) to top ($y = y_2$).

Observe, in (6), that the "inner" integral, $\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx$, is a function of y, say $F(y)$; the subsequently carried out "outer" integration is of the form $\int_{y_1}^{y_2} F(y)\,dy$, in which it is to be noted that the limits y_1 and y_2 are constants.

Alternatively, if we integrate first on y and then on x, as in Fig. 3b, we have

$$\int_{x_1}^{x_2}\left\{\int_{y_1(x)}^{y_2(x)} f(x,y)\,dy\right\}dx \qquad \text{or} \qquad \int_{x_1}^{x_2}\int_{y_1(x)}^{y_2(x)} f(x,y)\,dy\,dx. \tag{7}$$

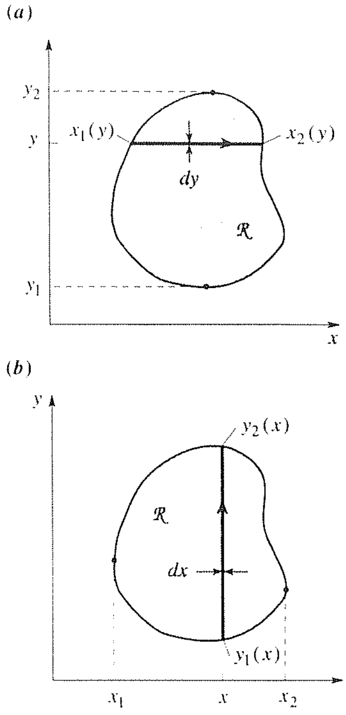


Figure 3. Iterated integrals.

Are the iterated integrals in (6) and (7) necessarily equal? This question is analogous to the one regarding *differentiation*: for example, are the mixed partial derivatives f_{xy} and f_{yx} equal? (We suggest that before continuing you review Theorem 13.3.1 and the Closure in Section 13.3, about interchanging the order of two limit processes.)

Returning to the question at hand, we state without proof that if $f(x,y)$ is continuous on $\mathcal{R}$, then all three quantities are equal: the double integral (1), the iterated integral (6), and the "inverted" iterated integral (7),*

*See J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Fransisco: W. H. Freeman, 1976), Chap.5, especially *Fubini's theorem* for rectangular regions on p.224 and the rectangularization idea on p.230. Note (above Example 1) that we have stated that all three integrals are equal if f is continuous on $\mathcal{R}$. If f is not continuous, it is possible for the iterated integrals in (6) and (7) to be unequal (or for one or both to fail to exist), and it is even possible for them to be equal, but not equal to the double integral $\iint_{\mathcal{R}} f(x,y)\,dA$. For such an example, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison–Wesley, 1957), Exer. 10-9.

$$\iint_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2}\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx\,dy = \int_{x_1}^{x_2}\int_{y_1(x)}^{y_2(x)} f(x,y)\,dy\,dx. \quad (8)$$

EXAMPLE 1. To illustrate the truth of (8), let us evaluate

$$I = \iint_{\mathcal{R}} xy^2\,dA, \quad (9)$$

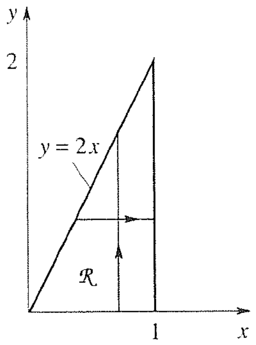


Figure 4. The region $\mathcal{R}$ in (9).

where $\mathcal{R}$ is the triangular region depicted in Fig. 4. First, observe that $f(x,y) = xy^2$ is indeed continuous on $\mathcal{R}$. If we integrate first on x, then $x_1(y) = y/2$, $x_2(y) = 1$, $y_1 = 0$, and $y_2 = 2$ so

$$\begin{aligned} I &= \int_0^2 \left\{ \int_{y/2}^1 xy^2\,dx \right\} dy = \int_0^2 \left\{ \frac{x^2y^2}{2}\bigg|_{y/2}^1 \right\} dy \\ &= \int_0^2 \left(\frac{y^2}{2} - \frac{y^4}{8} \right) dy = \frac{8}{6} - \frac{32}{40} = \frac{8}{15}. \end{aligned} \quad (10)$$

Alternatively, integrating first on y, as in (7), we have

$$I = \int_0^1 \left\{ \int_0^{2x} xy^2\,dy \right\} dx = \int_0^1 \left\{ \frac{xy^3}{3}\bigg|_0^{2x} \right\} dx = \frac{8}{15}, \quad (11)$$

which does agree with (10), in accordance with (8). ▮

EXAMPLE 2. Evaluate the iterated integral

$$I = \int_2^4 \int_{y/2}^{\sqrt{y}} e^{y/x}\,dx\,dy. \quad (12)$$

In Example 1 we were given a picture of $\mathcal{R}$, from which we deduced the integration limits $x_1(y) = y/2$, $x_2(y) = 1$, $y_1 = 0$, $y_2 = 0$ in (10). In the present example we are given the limits. Fine, but it is difficult to get started because $\exp(y/x)$ is a formidable function of x to integrate. It is more promising to invert the order of integration so we can integrate first with respect to y because $\exp(y/x)$ is a simple function of y. Thus, consider instead

$$I = \int_?^? \int_?^? e^{y/x}\,dy\,dx. \quad (13)$$

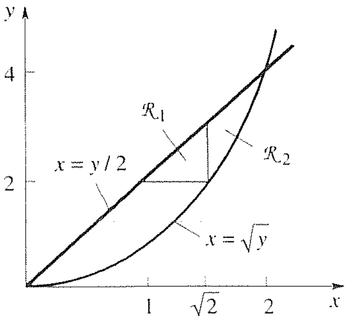


Figure 5. Region of integration.

But what are the new limits of integration? [They are *not* merely $\int_{y/2}^{\sqrt{y}} \int_2^4$, i.e., the reverse of the limits in (12).] To obtain the limits in (13) we need a picture of the region $\mathcal{R}$. That picture can be inferred from the limits in (12) and is the shaded region shown in Fig. 5. To integrate first on y and then on x we must break $\mathcal{R}$ into the subregions because the bottom boundaries of $\mathcal{R}_1$ and $\mathcal{R}_2$ are different, $y = 2$ for $\mathcal{R}_1$ and $y = x^2$ for $\mathcal{R}_2$. Thus,

$$I = \iint_{\mathcal{R}_1} e^{y/x}\,dy\,dx + \iint_{\mathcal{R}_2} e^{y/x}\,dy\,dx$$

$$= \int_1^{\sqrt{2}} \int_2^{2x} e^{y/x}\, dy\, dx + \int_{\sqrt{2}}^2 \int_{x^2}^{2x} e^{y/x}\, dy\, dx$$

$$= \int_1^{\sqrt{2}} \left\{ xe^{y/x} \Big|_2^{2x} \right\} dx + \int_{\sqrt{2}}^2 \left\{ xe^{y/x} \Big|_{x^2}^{2x} \right\} dx.$$

Thus,

$$I = \frac{e^2}{2} + (\sqrt{2} - 1)e^{\sqrt{2}} - \int_1^{\sqrt{2}} xe^{2/x}\, dx. \tag{14}$$

The remaining integral in (14) can be evaluated in terms of the tabulated exponential integral function, but let us not pursue that point. At the very least we have succeeded, by inverting the order of integration, in reducing I from a double integral to a single integral.

COMMENT. Be sure to see that the functional nature of the integrand was not relevant in our search for the new limits needed in (13). Rather, the original limits [in (12)] implied the region $\mathcal{R}$ (shaded in Fig. 6), and the region $\mathcal{R}$, in turn, implied the new limits. ▮

15.3.2. Triple integrals. Next, consider the triple integral

$$\iiint_{\mathcal{R}} f(x, y, z)\, dV. \tag{15}$$

We assume that $\mathcal{R}$ is a closed, bounded region in 3-space, that $f(x, y, z)$ is defined in $\mathcal{R}$, and that the boundary $\mathcal{S}$ of $\mathcal{R}$ is a piecewise smooth surface. By a **smooth surface** we mean one that has a continuously turning normal,* and by a **piecewise smooth surface** we mean one that consists of at most a finite number of smooth surfaces arranged edge to edge.

To define the triple integral (15), we partition $\mathcal{R}$ into a cellular structure analogous to that shown in Fig. 1. This time, however, the cells are rectangular prisms rather than rectangles. Corresponding to this N-cell partition we form the Riemann sum

$$\sum_{j=1}^{N} f(x_j, y_j, z_j) \Delta V_j, \tag{16}$$

where $\Delta V_j = \Delta x_j \Delta y_j \Delta z_j$ is the volume of the jth cell. Then we define a finer partition, compute its Riemann sum, and repeat the procedure over and over (in principle; one does not actually carry out this procedure in practice). If the sequence of values of the Riemann sum thus generated converges to a limit that limit is, by definition, the **triple integral** (15). That is,

$$\iiint_{\mathcal{R}} f(x, y, z)\, dV \equiv \lim_{|p| \to 0} \sum f(x_j, y_j, z_j) \Delta V_j \tag{17}$$

*For present purposes, one's intuitive notion of the normal to a surface should suffice; the term is defined in Section 15.4.

if the limit on the right exists, where $|p|$ is the *norm* of the partition, namely, the greatest of the dimensions $\Delta x_j, \Delta y_j, \Delta z_j$ for $j = 1, \dots, N$. It can be shown that the limit (17) will indeed exist if f is continuous in the closed region $\mathcal{R}$, that condition being sufficient but not necessary.

We state, without writing them out, that properties (4a) to (4c) carry over in the obvious way to the case of triple integrals. For example, the **mean value theorem** for triple integrals states that if $f(x, y, z)$ is continuous throughout the closed region $\mathcal{R}$, then there exists at least one point (x_0, y_0, z_0) in $\mathcal{R}$ such that

$$\boxed{\iiint\limits_{\mathcal{R}} f(x,y,z)\,dV = f(x_0, y_0, z_0)V,} \tag{18}$$

where

$$V = \iiint\limits_{\mathcal{R}} dV \tag{19}$$

is the **volume** of $\mathcal{R}$.

To evaluate a triple integral analytically, we treat it as an iterated integral and integrate on one variable at a time.

EXAMPLE 3. Evaluate

$$I = \iiint\limits_{\mathcal{R}} yz^2\,dV, \tag{20}$$

where $\mathcal{R}$ is the closed region bounded by the planes $x = 0$, $y = 0$, $z = 0$, and $3x + 2y + 6z = 6$. With $dV = dx\,dy\,dz$, we write

$$I = \int_?^? \int_?^? \int_?^? yz^2\,dx\,dy\,dz, \tag{21}$$

but how do we obtain the integration limits? It is helpful to re-express (21) as

$$I = \int_?^? \left\{ \int_?^? \int_?^? yz^2\,dx\,dy \right\} dz \tag{22}$$

and to attend first to the bracketed iterated integral on x and y. In that iterated integral z is held fixed, at some value between $z = 0$ and $z = 1$ because we are not yet integrating on z. Thus, the x, y integration is being carried out on a horizontal slice such as the triangle cde (Fig. 6). In turn, within that iterated integral we are integrating on x first, with y held fixed at some point between c and e. In other words, first we integrate on x from a ($x = 0$) to b ($x = 2 - \frac{2}{3}y - 2z$), then we sweep the line ab from cd ($y = 0$) to e ($y = 3 - 3z$). Finally, we sweep the triangle cde from $z = 0$ to $z = 1$. Thus, we have

$$I = \int_0^1 \int_0^{3-3z} \int_0^{2-\frac{2}{3}y-2z} yz^2\,dx\,dy\,dz$$

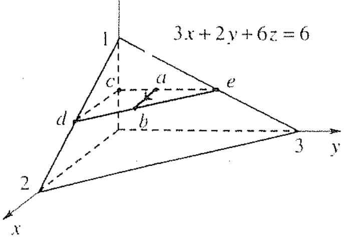


Figure 6. The region $\mathcal{R}$.

$$= \int_0^1 \int_0^{3-3z} yz^2 x \Big|_{x=0}^{x=2-\frac{2}{3}y-2z} dy\,dz$$
$$= \int_0^1 \int_0^{3-3z} yz^2 \left(2 - \frac{2}{3}y - 2z\right) dy\,dz$$
$$= \int_0^1 z^2 \left(y^2 - \frac{2}{9}y^3 - y^2 z\right) \Bigg|_{y=0}^{y=3-3z} dz$$
$$= \int_0^1 z^2 \left[9(1-z)^2 - 6(1-z)^3 - 9(1-z)^2 z\right] dz = \frac{1}{20}. \tag{23}$$

In case the upper y limit ($y = 3 - 3z$) is not clear, observe that the point e is on the plane $3x + 2y + 6z = 6$ at $x = 0$, hence $y + 3z = 3$ and $y = 3 - 3z$.

COMMENT 1. The limits are *not* $\int_0^1 \int_0^3 \int_0^2$. The latter would be correct if $\mathcal{R}$ were a rectangular prism, but it is not. The key to determining the correct limits lies in carrying out the sketch (Fig. 6), including the triangle cde and the line ab.

COMMENT 2. How do we know that x varies from 0 (at a) to $2 - \frac{2}{3}y - 2z$ (at b) and not vice versa, and similarly for the y and z limits? The key is that $dV = dx\,dy\,dz$ must be positive because it is a volume element. We can ensure that each dV is positive by choosing the limits as we have, so that each dx, dy, and dz is positive; for instance, if the z limits had been chosen as 1 to 0, instead of 0 to 1, then each dz would be *negative*.

COMMENT 3. As for double integrals, if the integrand is continuous on the region $\mathcal{R}$, then the triple integral is equal to the iterated integral, with the iteration carried out in any order. That is,

$$\iiint_{\mathcal{R}} f\,dV = \int_{z_1}^{z_2} \int_{y_1(z)}^{y_2(z)} \int_{x_1(y,z)}^{x_2(y,z)} f\,dx\,dy\,dz$$
$$= \int_{z_1}^{z_2} \int_{x_1(z)}^{x_2(z)} \int_{y_1(x,z)}^{y_2(x,z)} f\,dy\,dx\,dz = \cdots, \tag{24}$$

and so on, where the three dots are included because there are actually six possible integration sequences: xyz, yxz, zxy, xzy, zyx, and yzx. Integrating in the sequence zyx, for instance, we would have

$$I = \int_0^2 \int_0^{3-\frac{3}{2}x} \int_0^{1-\frac{1}{2}x-\frac{1}{3}y} yz^2\,dz\,dy\,dx, \tag{25}$$

where the limits are obtained from Fig. 6. Evaluation of the iterated integral in (25) once again gives $I = 1/20$, as in (23).

COMMENT 4. Observe the pattern in (24); for instance, consider the final iterated integral, in which we integrate first on y, then on x, and then on z. The limits can depend only on the variables that have not yet been integrated. Thus, the y limits can, in general, depend on x and z, the x limits can depend only on z, and the z limits must be constants. ▮

Closure. This section is included as a brief review of double and triple integrals. Of special importance are the properties (4a)–(4c) and the corresponding ones for triple integrals, the distinction between double and triple integrals and their iterated integral counterparts, and the determination of integration limits.

Computer software. Besides single integrals, the *Maple* int command can be used to evaluate multiple integrals. For instance, to evaluate the integral (9), enter

$$\text{int}(\text{int}(x * y\hat{}2,\ x = y/2..1),\ y = 0..2);$$

and return. The result is 8/15, as obtained analytically in Example 1.

EXERCISES 15.3

Exercises 1–9 correspond to Section 15.3.1, and the remainder correspond to Section 15.3.2.

1. Evaluate each of the following iterated integrals. Then integrate again, with the order of integration reversed, and show that the same result is obtained.

(a) $\int_0^1 \int_0^y y^2\,dx\,dy$ (b) $\int_0^1 \int_x^{2x} x^3 y\,dy\,dx$

(c) $\int_0^2 \int_y^2 xe^{x+y}\,dx\,dy$ (d) $\int_1^2 \int_0^y \sin(x-y)\,dx\,dy$

(e) $\int_0^1 \int_y^{(2y+1)/4} (x+y)^8\,dx\,dy$ (f) $\int_0^1 \int_{-y}^{2-y} x\,dx\,dy$

(g) $\int_{-1}^3 \int_{-x}^x x^2\,dy\,dx$ (h) $\int_1^4 \int_2^{\sqrt{x}} x^2\,dy\,dx$

2. Show that $\int_0^T \int_0^t x(\tau)\,d\tau\,dt$ can be reduced to the single integral $\int_0^T (T-\tau)x(\tau)\,d\tau$ (T = constant). HINT: See the Comment at the end of Example 2.

3. Show that

(a) $\int_0^6 \int_{y/2}^3 \frac{1}{x} e^{y/x}\,dx\,dy = 3(e^2 - 1)$

(b) $\int_4^{16} \int_{y/2}^{\sqrt{y}} \frac{1}{x} e^{y/x}\,dx\,dy = 7e^2 - e^4 - \int_4^8 e^{16/x}\,dx$

(c) $\int_0^1 \int_{y/2}^y \sin(x^2 y)\,dx\,dy = \int_0^{1/2} \frac{\cos(x^3) - \cos(2x^3)}{x^2}\,dx$
$+ \int_{1/2}^1 \frac{\cos(x^3) - \cos(x^2)}{x^2}\,dx$

(d) $\int_0^{\pi/2} \int_x^{\pi/2} \frac{\sin y}{y}\,dy\,dx = 1$

4. (*Mass and center of gravity*) Let $\sigma(x,y)$ be the density of a distribution of mass over a region $\mathcal{R}$ in the x,y plane [i.e., $\sigma(x,y)$ is the mass per unit area at (x,y)]. Then the x,y coordinates of the **center of gravity** are defined as

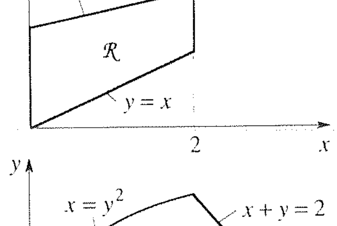
$$x_c = \frac{1}{M}\iint_{\mathcal{R}} x\sigma(x,y)\,dA,$$
$$y_c = \frac{1}{M}\iint_{\mathcal{R}} y\sigma(x,y)\,dA, \tag{4.1}$$

where

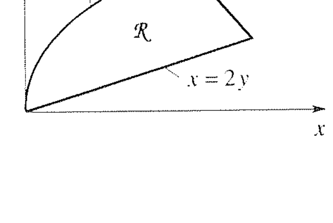
$$M = \iint_{\mathcal{R}} \sigma(x,y)\,dA \tag{4.2}$$

is the total mass. Evaluate x_c and y_c in each case. In parts (a) to (d), $\sigma(x,y)$ is a constant, say σ.

(a)

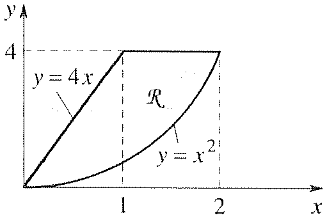


(b)



(c)

(d)



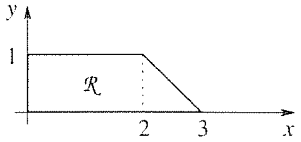
(e) Same $\mathcal{R}$ as in part (a), but $\sigma(x,y) = 1 + x$
(f) Same $\mathcal{R}$ as in part (a), but $\sigma(x,y) = y$
(g) Same $\mathcal{R}$ as in part (c), but $\sigma(x,y) = x$
(h) Same $\mathcal{R}$ as in part (c), but $\sigma(x,y) = y$

5. (*Moments of inertia*) Let $\sigma(x,y)$ be the density of a distribution of mass over a region $\mathcal{R}$ in the x,y plane; i.e., $\sigma(x,y)$ is the mass per unit area at (x,y). Then the **moments of inertia** I_x and I_y, about the x,y axes, respectively, are defined as

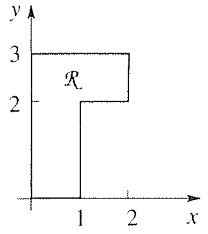
$$I_x = \iint_{\mathcal{R}} y^2\sigma(x,y)\,dA, \qquad I_y = \iint_{\mathcal{R}} x^2\sigma(x,y)\,dA. \tag{5.1}$$

Evaluate I_x and I_y in each case. In parts (a) and (b), $\sigma(x,y)$ is a constant, say σ.

(a)



(b)



(c) $\mathcal{R}$, σ as in Exercise 4(a)
(d) $\mathcal{R}$, σ as in Exercise 4(b)
(e) $\mathcal{R}$, σ as in Exercise 4(c)
(f) $\mathcal{R}$, σ as in Exercise 4(d)
(g) $\mathcal{R}$, σ as in Exercise 4(e)
(h) $\mathcal{R}$, σ as in Exercise 4(f)

6. The **exponential integral** function $E_1(t)$ is defined by the formula

$$\boxed{E_1(t) = \int_t^\infty \frac{e^{-\tau}}{\tau}\,d\tau, \quad (t > 0)} \tag{6.1}$$

and is tabulated in the literature, just as are $\sin t, e^t$, and so on. Evaluate the integral $\int_0^x E_1(t)\,dt$, and show that it equals $1 - e^{-x} + xE_1(x)$. (See the index for applications of the exponential integral function within this text.)

7. If

$$f(x,y) = \begin{cases} 1, & 0 \le y \le 2x, \quad 0 \le x \le 1 \\ 0, & \text{elsewhere,} \end{cases}$$

show that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{u} f(x,y)\,dx\,dy = \begin{cases} 0, & u \le 0 \\ u^2, & 0 < u < 1 \\ 1, & u \ge 1. \end{cases}$$

8. Use computer software to evaluate the integral in (14), and thus obtain a numerical value for I.

9. (a) – (h) Use computer software to evaluate the iterated integral given in the corresponding part of Exercise 1.

10. Evaluate each of the following iterated integrals.

(a) $\int_1^2\int_0^{3x}\int_{2y}^{x} dz\,dy\,dx$
(b) $\int_{-1}^1\int_0^z\int_2^3 z^2\sin(yz)\,dx\,dy\,dz$
(c) $\int_0^\pi\int_0^{2y}\int_0^{y+z}\cos(x+y)\,dx\,dz\,dy$
(d) $\int_{-1}^3\int_0^{z^2}\int_{\pi y}^0 \sin\dfrac{x}{y}\,dx\,dy\,dz$
(e) $\int_0^2\int_0^{4z}\int_0^{2-z} e^{x+y}\,dx\,dy\,dz$
(f) $\int_0^1\int_0^1\int_0^y z^3y^2e^{xyz}\,dx\,dy\,dz$

11. Evaluate

$$I = \int_0^1\int_0^{\pi z}\int_{y/\pi}^{z} \sin\frac{y}{x}\,dx\,dy\,dz.$$

HINT: The x integration looks quite difficult, so try inverting the order of integration to the form

$$I = \int_0^1\int_?^?\int_?^? \sin\frac{y}{x}\,dy\,dx\,dz.$$

To determine the "?" integration limits we do *not* need to draw the full three-dimensional region of integration. Rather, a look at the x,y plane (with z regarded as a *fixed* value, somewhere between its limits of 0 and 1) will suffice since it is only the x and y integrations that we are interchanging.

12. Evaluate the following integrals. The idea, in each case, is the same as discussed in Exercise 4.

(a) $\mathcal{R}$: the tetrahedron with vertices at $(1,0,0)$, $(0,2,0)$, $(0,0,1)$, $(0,0,0)$; $\sigma =$ constant. Evaluate x_c.

(b) $\mathcal{R}$: the tetrahedron with vertices at $(1,0,1)$, $(0,0,1)$, $(0,1,1)$, $(0,0,0)$; $\sigma = x$. Evaluate z_c.
(c) $\mathcal{R}$: the finite region bounded by $z = 4 - x^2 - y^2$, and the x, y plane; $\sigma =$ constant. Evaluate z_c.
(d) $\mathcal{R}$: the finite region bounded by $z = 1 - x^2 - y^2$, the x, z plane, and the x, y plane; $\sigma =$ constant. Evaluate z_c.
(e) Repeat part (d), but with $\sigma = z$.

13. (*Moments of inertia*) Let $\sigma(x, y, z)$ be the density of a distribution of a mass over a region $\mathcal{R}$ in x, y, z space [i.e., $\sigma(x, y, z)$ is the mass per unit volume at (x, y, z)]. Then the **moments of inertia** I_x, I_y, and I_z, about the x, y, z axes, respectively, are defined as

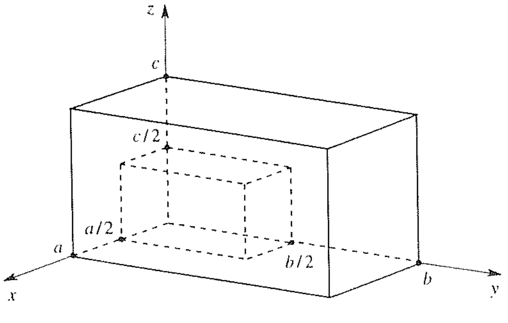
$$I_x = \iiint_{\mathcal{R}} (y^2 + z^2)\, \sigma(x, y, z)\, dV$$

$$I_y = \iiint_{\mathcal{R}} (x^2 + z^2)\, \sigma(x, y, z)\, dV$$

$$I_z = \iiint_{\mathcal{R}} (x^2 + y^2)\, \sigma(x, y, z)\, dV.$$

In each case evaluate I_x.

(a) $\mathcal{R}$, σ as given in Exercise 12(a).
(b) $\mathcal{R}$ the rectangular prism $0 < x < a$, $0 < y < b$, $0 < z < c$; $\sigma =$ constant. Show that $I_x = \sigma abc(b^2 + c^2)/3$.
(c) $\mathcal{R}$ the rectangular prism $0 < x < a$, $0 < y < b$, $0 < z < c$, but with the corner $0 < x < a/2$, $0 < y < b/2$, $0 < z < c/2$ cut out, as shown in the figure; $\sigma =$ constant. HINT: You may use the result given in part (b).



14. Determine the "?" integration limits. HINT: You can use the given limits to infer the region $\mathcal{R}$. Sketching $\mathcal{R}$, you can then determine the new limits. However, that method might involve a challenging three-dimensional sketch and would not work at all for a quadruple integral, say, because $\mathcal{R}$ would then be four-dimensional. Thus, we suggest that you use one or more two-dimensional pictures. In part (a), for instance, only the y and z integrations are being interchanged so it suffices to consider the y, z plane. In part (b) you can accomplish the desired change in the order of integration by using two successive interchanges, as follows: $xyz \to yxz \to yzx$. The first interchange requires a look at the x, y plane only, and the second interchange requires a look at the x, z plane only.

(a) $$\int_0^1 \int_0^z \int_0^z f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dx\, dz\, dy$$

(b) $$\int_0^1 \int_0^z \int_0^{2z} f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dy\, dz\, dx$$

(c) $$\int_0^1 \int_0^z \int_0^z f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dz\, dy\, dx$$

(d) $$\int_0^2 \int_0^z \int_z^2 f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dy\, dz\, dx$$

(e) $$\int_0^4 \int_{z/2}^{\sqrt{z}} \int_0^{y+z} f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dx\, dz\, dy$$

(f) $$\int_4^6 \int_1^3 \int_0^2 f(x, y, z)\, dx\, dy\, dz = \int_?^? \int_?^? \int_?^? f(x, y, z)\, dz\, dy\, dx$$

15. (a)–(f) Use computer software to evaluate the integral in the corresponding part of Exercise 10.

15.4 Surfaces

Aiming at a consideration of surface integrals in Section 15.5, we begin by discussing the parametric representation of surfaces, the tangent plane, and the normal.

15.4.1. Parametric representation of surfaces. To begin, recall that a space curve may be represented parametrically by equations

$$x = x(u), \quad y = y(u), \quad z = z(u) \tag{1}$$

over some interval of the parameter u. (Previously we used τ as our parameter; here we prefer u.) Or, in vector form, we express the position vector $\mathbf{R}(u)$, from the origin of the reference x, y, z coordinate system to a point on the curve, as

$$\mathbf{R}(u) = x(u)\hat{\mathbf{i}} + y(u)\hat{\mathbf{j}} + z(u)\hat{\mathbf{k}}. \tag{2}$$

Similarly, it is natural to expect a *two*-parameter family of curves

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v) \tag{3}$$

or

$$\mathbf{R}(u, v) = x(u, v)\hat{\mathbf{i}} + y(u, v)\hat{\mathbf{j}} + z(u, v)\hat{\mathbf{k}}, \tag{4}$$

over some intervals of the real parameters u and v, to define a surface. That is, for each fixed v the parametrization (3) [or (4)] defines a curve (in general). Thus, as we vary v we produce a family of such curves which, in general, will generate a surface.

EXAMPLE 1. Let

$$\begin{aligned} x &= a \sin v \cos u, \\ y &= a \sin v \sin u, \\ z &= a \cos v, \end{aligned} \tag{5}$$

where $0 \leq u \leq \pi/2$, $0 \leq v \leq \pi/2$, and a is a positive constant. Squaring and adding these three equations gives $x^2 + y^2 + z^2 = a^2$, which represents a spherical surface of radius a, centered at the origin. In fact, a comparison of (5) with (26a) in Section 14.6 reveals that a, v, u are actually the spherical coordinates ρ, ϕ, θ, respectively, in disguise. Since $\rho = a =$ constant, the surface is spherical. More precisely, it is one-eighth of a spherical surface as shown in Fig. 1b, where it is denoted as $\mathcal{S}$. Thus, in this example the parameters u and v happen to bear a geometrical significance; they are the spherical polar angles shown in Fig. 1b.

If v is fixed and u is varied over the line AB in Fig. 1a, then (5) generates the curve $A'B'$ shown in Fig. 1b. Similarly, if u is fixed and v is varied over CD, then (5) generates the curve $C'D'$. If we think of $\mathcal{S}$ as part of the earth's globe, with C' as the north pole, then the $v =$ constant curves are the lines of latitude and the $u =$ constant curves are the lines of longitude.

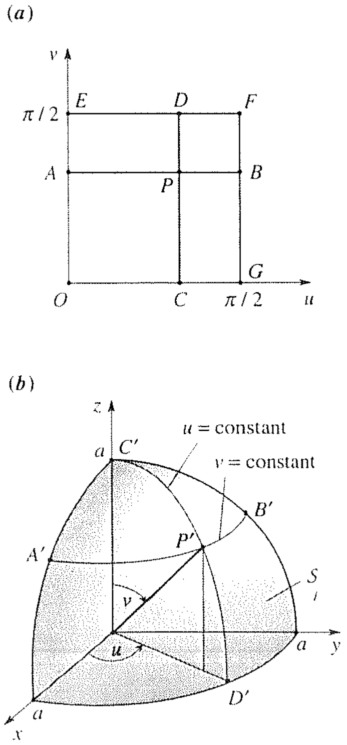


Figure 1. The surface $\mathcal{S}$ generated by (5).

COMMENT 1. It is common to refer to u, v as **curvilinear coordinates** of the generated surface $\mathcal{S}$. In this example the network of lines of latitude (the $v =$ constant curves) and longitude (the $u =$ constant curves) forms a coordinate system on the surface of the sphere just as a rectilinear network of $x =$ constant and $y =$ constant lines forms a coordinate system on the x, y plane. We call the $u =$ constant and $v =$ constant curves on $\mathcal{S}$ the **coordinate curves**. In particular, along the $v =$ constant curve only u is changing so we call the $v =$ constant curves the **u coordinate curves**. Similarly, we call the $u =$ constant curves the **v coordinate curves**. For instance, in Fig. 1b $A'B'$ is a u coordinate curve, with u increasing from 0 at A' to $\pi/2$ at B', and $C'D'$ is a v coordinate curve, with v increasing from 0 at C' to $\pi/2$ at D'. In Fig. 2 we show many such u and v coordinate curves so as to more clearly see the u, v coordinate "mesh" on the spherical surface.

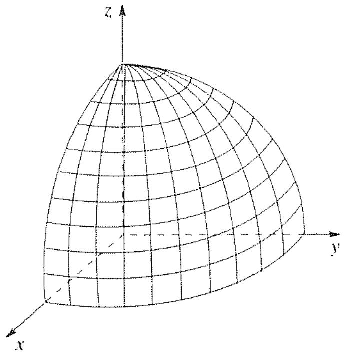


Figure 2. u, v coordinate mesh on $\mathcal{S}$.

COMMENT 2. If the surface $\mathcal{S}$ can be expressed in the form $z = f(x, y)$, then we can parametrize $\mathcal{S}$ by $x = u$, $y = v$, and $z = f(u, v)$. In the present case,

$$z = \sqrt{a^2 - x^2 - y^2} \tag{6}$$

so an alternative parametrization, to that given by (5), is

$$x = u, \qquad y = v, \qquad z = \sqrt{a^2 - u^2 - v^2}. \tag{7}$$

In that case the u, v coordinate curves through a given point P on $\mathcal{S}$ are as shown in Fig. 3a, and the overall coordinate mesh is as shown in Fig. 3b. ▮

15.4.2. Tangent plane and normal. Let a given vector function $\mathbf{R}(u, v)$, for some u and v intervals, define a surface $\mathcal{S}$. Since v is held constant in the partial derivative $\mathbf{R}_u$, it follows that $\mathbf{R}_u$ (if it exists and is nonzero) is a tangent vector to $\mathcal{S}$, along the u coordinate curve (and in the increasing u direction) as shown in Fig. 4. Similarly, $\mathbf{R}_v$ (if it exists and is nonzero) is a tangent vector to $\mathcal{S}$ along the v coordinate curve (and in the increasing v direction).

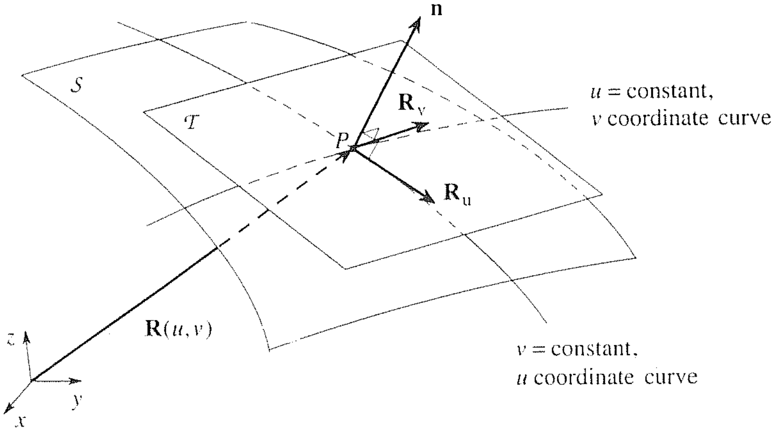


Figure 4. Tangent plane and normal.

From Fig. 4 we see that a **normal** vector to $\mathcal{S}$, at P, can be obtained from $\mathbf{R}(u, v)$ as

$$\mathbf{n} = \mathbf{R}_u \times \mathbf{R}_v \tag{8}$$

or

$$\boxed{\hat{\mathbf{n}} = \frac{\mathbf{R}_u \times \mathbf{R}_v}{\|\mathbf{R}_u \times \mathbf{R}_v\|},} \tag{9}$$

provided that

$$\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0} \tag{10}$$

at P, that is, provided that $\mathbf{R}_u$ and $\mathbf{R}_v$ are nonzero and noncollinear (i.e., provided that $\mathbf{R}_u$ and $\mathbf{R}_v$ are linearly independent). In Example 1, $\mathbf{R}_u$ and $\mathbf{R}_v$ happen to be not only linearly independent but even orthogonal at each point on $\mathcal{S}$, but in general the u and v coordinate curves need not be orthogonal.

If (10) is satisfied, then $\mathbf{R}_u$ and $\mathbf{R}_v$ determine a plane, the **tangent plane** $\mathcal{T}$ to $\mathcal{S}$ at P. For if $u = u(\tau)$, $v = v(\tau)$ are parametric equations of any curve $\mathcal{C}$ through P, in $\mathcal{S}$, then $d\mathbf{R}/d\tau$ is tangent to $\mathcal{C}$ at P. But

$$\frac{d\mathbf{R}}{d\tau} = \frac{d}{d\tau}\mathbf{R}(u(\tau), v(\tau)) = \mathbf{R}_u u' + \mathbf{R}_v v' \tag{11}$$

is in $\mathcal{T}$ because it is a linear combination of $\mathbf{R}_u$ and $\mathbf{R}_v$. (For the second equality, which is chain differentiation, to hold, let us assume that $\mathbf{R}_u$ and $\mathbf{R}_v$ are continuous functions of u and v at P and, of course, that u and v are differentiable functions of τ at that point.)

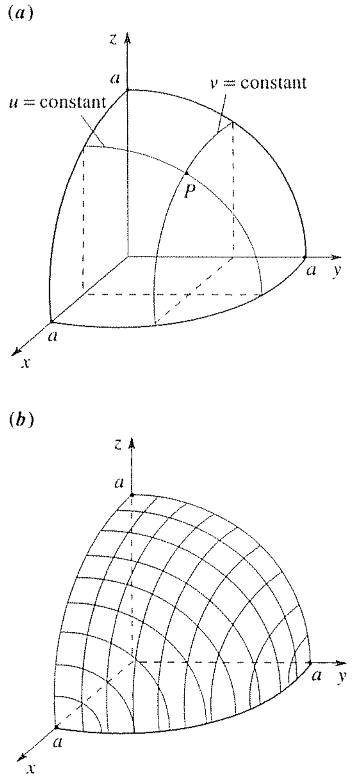


Figure 3. Using (7), instead.

Given $\mathbf{R}(u, v)$, and a point P on $\mathcal{S}$, how can we obtain the equation of the tangent plane $\mathcal{T}$ to $\mathcal{S}$ at P? Let $\mathbf{R}_p = x_p\hat{\mathbf{i}} + y_p\hat{\mathbf{j}} + z_p\hat{\mathbf{k}}$ and $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$ be position vectors to P, and to any point on $\mathcal{T}$, respectively (Fig. 5). Then

$$(\mathbf{R} - \mathbf{R}_p) \cdot \hat{\mathbf{n}} = 0, \tag{12a}$$

where $\hat{\mathbf{n}}$ can be computed from (9) and is a known vector, as is $\mathbf{R}_p$. If we denote $\hat{\mathbf{n}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$, say, then (12a) becomes

$$\boxed{a(x - x_p) + b(y - y_p) + c(z - z_p) = 0} \tag{12b}$$

or

$$\boxed{ax + by + cz = d,} \tag{12c}$$

where $d = ax_p + by_p + cz_p$. Equations (12) are equivalent, each one being the equation of the tangent plane at P; (12b) and (12c) may well look familiar from the calculus.

Naturally, the normal $\hat{\mathbf{n}}$ can be determined only to within a factor of ± 1. For instance, if $\hat{\mathbf{n}} = (\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}})/\sqrt{6}$ is a normal to $\mathcal{S}$ at some particular point P, then so is $\hat{\mathbf{n}} = -(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}})/\sqrt{6}$. The two possible $\hat{\mathbf{n}}$ vectors at P determine a line which we call the **normal line** at P. A unique tangent plane exists at P if and only if a unique normal line exists there, and the condition (10) is *sufficient, but not necessary* (see Exercise 10), for the existence of a unique normal line.

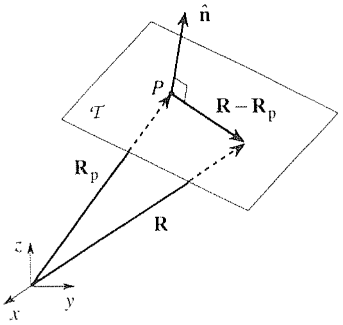


Figure 5. Obtaining equation of tangent plane.

EXAMPLE 2. For the surface $\mathcal{S}$ defined parametrically by

$$x = u + v,$$

$$\begin{aligned} y &= u - v^2, \\ z &= u^2 + 3v^4, \end{aligned} \tag{13}$$

find the tangent plane to $\mathcal{S}$ at $u = 0$, $v = 1$, that is, at the point $(x, y, z) = (1, -1, 3)$. Since $\mathbf{R} = (u + v)\hat{\mathbf{i}} + (u - v^2)\hat{\mathbf{j}} + (u^2 + 3v^4)\hat{\mathbf{k}}$, we have

$$\begin{aligned} \mathbf{R}_u &= \hat{\mathbf{i}} + \hat{\mathbf{j}} + 2u\hat{\mathbf{k}} = \hat{\mathbf{i}} + \hat{\mathbf{j}}, \\ \mathbf{R}_v &= \hat{\mathbf{i}} - 2v\hat{\mathbf{j}} + 12v^3\hat{\mathbf{k}} = \hat{\mathbf{i}} - 2\hat{\mathbf{j}} + 12\hat{\mathbf{k}} \end{aligned}$$

at $u = 0$ and $v = 1$ so (9) gives

$$\hat{\mathbf{n}} = \frac{4\hat{\mathbf{i}} - 4\hat{\mathbf{j}} - \hat{\mathbf{k}}}{\sqrt{33}}. \tag{14}$$

Thus, (12b) gives

$$\frac{4}{\sqrt{33}}(x - 1) - \frac{4}{\sqrt{33}}(y + 1) - \frac{1}{\sqrt{33}}(z - 3) = 0 \tag{15}$$

or

$$4x - 4y - z = 5 \tag{16}$$

as the equation of the desired tangent plane. ▮

EXAMPLE 3. The relation

$$x^2 + y^2 + z^2 + z^4 = 1 \tag{17}$$

defines a surface that is somewhat of a sphere compressed at its north and south poles; its intercepts along the coordinate axes are ±1 along each of the x and y axes, and ±0.618 along the z axis. Find the equation of the tangent plane at the point $(x, y, z) = \left(\frac{1}{2}, \frac{1}{2}, \sqrt{\frac{\sqrt{3}-1}{2}}\right)$ $= (0.5, 0.5, 0.605)$ on $\mathcal{S}$.

To use (12b), say, we need to know a, b, c; that is, we need to know $\hat{\mathbf{n}}$. Since $\hat{\mathbf{n}}$ is given, by (9), in terms of $\mathbf{R}(u, v)$, our first step is to parametrize $\mathcal{S}$ in order to find $\mathbf{R}(u, v)$. We can solve (17) (with the help of the quadratic formula) for z, in the form $z = f(x, y)$, and then parametrize $\mathcal{S}$ by $x = u$, $y = v$, $z = f(u, v)$, but it is easier to solve (17) for x as

$$x = \sqrt{1 - y^2 - z^2 - z^4}, \tag{18}$$

and hence to parametrize $\mathcal{S}$ by $y = u$, $z = v$, $x = \sqrt{1 - u^2 - v^2 - v^4}$. Then

$$\mathbf{R}(u, v) = \sqrt{1 - u^2 - v^2 - v^4}\,\hat{\mathbf{i}} + u\hat{\mathbf{j}} + v\hat{\mathbf{k}} \tag{19}$$

so, with $u = 0.5$ and $v = 0.605$, (19) gives $\mathbf{R}_u = -\hat{\mathbf{i}} + \hat{\mathbf{j}}$ and $\mathbf{R}_v = -2.096\hat{\mathbf{i}} + \hat{\mathbf{k}}$. Next, (9) gives

$$\hat{\mathbf{n}} = 0.396\hat{\mathbf{i}} + 0.396\hat{\mathbf{j}} + 0.829\hat{\mathbf{k}}. \tag{20}$$

Hence, $a = 0.396$, $b = 0.396$, $c = 0.829$, $x_p = 0.5$, $y_p = 0.5$, and $z_p = 0.605$ so (12b) gives

$$0.396x + 0.396y + 0.829z = 0.898 \tag{21}$$

as the equation of the desired tangent plane.

COMMENT. Alternatively, if $\mathcal{S}$ is given in the form of a relation

$$f(x, y, z) = 0, \tag{22}$$

then the Taylor series expansion about (x_p, y_p, z_p) gives

$$\begin{aligned} f(x_p, y_p, z_p) + f_x(x_p, y_p, z_p)(x - x_p) + f_y(x_p, y_p, z_p)(y - y_p) \\ + f_z(x_p, y_p, z_p)(z - z_p) + \cdots = 0, \end{aligned} \tag{23}$$

where the three dots denote the terms of second order and higher. If we linearize by dropping those higher-order terms, and note that $f(x_p, y_p, z_p) = 0$ [because $f(x, y, z) = 0$ is the equation of $\mathcal{S}$, and (x_p, y_p, z_p) is on $\mathcal{S}$], then (23) becomes

$$\boxed{f_x(x_p, y_p, z_p)(x - x_p) + f_y(x_p, y_p, z_p)(y - y_p) + f_z(x_p, y_p, z_p)(z - z_p) = 0,} \tag{24}$$

which is the same as (12b), but expressed in terms of $f(x, y, z)$ rather than in terms of $\mathbf{R}(u, v)$. We leave it as an exercise to show that, with $f(x, y, z) = x^2 + y^2 + z^2 + z^4 - 1$, (24) agrees with (21). ▮

Closure. A surface $\mathcal{S}$ may be represented by the parametric equations $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ or, equivalently, by the position vector

$$\mathbf{R}(u, v) = x(u, v)\hat{\mathbf{i}} + y(u, v)\hat{\mathbf{j}} + z(u, v)\hat{\mathbf{k}}.$$

Accordingly, $\mathcal{S}$ is covered by a mesh of u and v coordinate curves (namely, the $v =$ constant and $u =$ constant curves, respectively), as shown in Figs. 2 and 3b for two different parametrizations. From $\mathbf{R}(u, v)$ we can obtain the normal to $\mathcal{S}$ from (9) (except at points on $\mathcal{S}$ at which $\mathbf{R}_u \times \mathbf{R}_v = \mathbf{0}$) and, knowing $\hat{\mathbf{n}}$, the condition (12a) gives us the equation of the tangent plane in the forms (12b) or (12c). Alternatively, if in place of $\mathbf{R}(u, v)$ we know $\mathcal{S}$ through a relation $f(x, y, z) = 0$, then Taylor-expanding f and linearizing gives the equation of the tangent plane in the form (24). Naturally, from a comparison of (12) and (24) we can see that the normal vector to $\mathcal{S}$ can be expressed, in terms of f, as

$$\hat{\mathbf{n}} = \frac{f_x\hat{\mathbf{i}} + f_y\hat{\mathbf{j}} + f_z\hat{\mathbf{k}}}{\sqrt{f_x^2 + f_y^2 + f_z^2}}, \tag{25}$$

where f_x, f_y, f_z are evaluated at the point (x_p, y_p, z_p) on $\mathcal{S}$.

EXERCISES 15.4

1. In Fig. 1 we show the curves on $\mathcal{S}$ that are the images of the lines AB and CD in the u, v plane, namely, $A'B'$ and $C'D'$, respectively. Sketch, trace, or photocopy the surface $\mathcal{S}$ in Fig. 1b and show the curve on $\mathcal{S}$ that is the image of the following line in the u, v plane.

(a) OE (b) EF (c) FG (d) OG
(e) the line $v = u$ from O to F
(f) the line $v = \pi/2 - u$ from E to G

2. Consider the parametrization

$$x = u, \quad y = u + v, \quad z = 0,$$

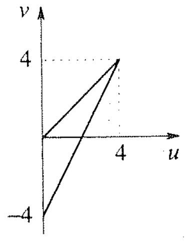
over $0 < u < 4$, $0 < v < 4$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = 1, 2, 3$ and $v = 1, 2, 3$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Are there any points on $\mathcal{S}$ at which (10) is not satisfied?

3. Consider the parametrization

$$x = u, \quad y = u - v, \quad z = 4 - 2u + v, \tag{3.1}$$

over the open region shown (shaded) below.



(a) Eliminating u and v from (3.1), obtain a nonparametric representation of the surface $\mathcal{S}$ [i.e., in the form $f(x, y, z) = 0$], and show $\mathcal{S}$ in a neat, labeled sketch.
(b) Show the $u = 1, 2, 3$ and $v = 1, 2, 3$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9).

4. Consider the parametrization

$$x = u \cos v, \quad y = u \sin v, \quad z = 0,$$

over $0 \le u < 3$, $0 \le v < 2\pi$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = 1, 2$ and $v = 0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that (10) is *not* satisfied at the origin, but that a unit normal $\hat{\mathbf{n}}$ does, nevertheless, exist there.

5. Consider the parametrization

$$x = au \cos v, \quad y = bu \sin v, \quad z = 0,$$

over $0 \le u < 1$, $0 \le v < 2\pi$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = \frac{1}{2}$ and $v = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that (10) is *not* satisfied at the origin, but that a unit normal $\hat{\mathbf{n}}$ does, nevertheless, exist there.

6. Consider the parametrization

$$x = (1 - u) \cos v, \quad y = (1 - u) \sin v, \quad z = u, \tag{6.1}$$

over $0 < u \le 1$, $0 \le v < 2\pi$.

(a) Eliminating u and v from (6.1), obtain a nonparametric representation of the surface $\mathcal{S}$, and show $\mathcal{S}$ in a neat, labeled sketch. Identify $\mathcal{S}$. (For example, is it a plane? A sphere? . . .)
(b) Show the $u = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ and $v = 0, \pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that there is one point on $\mathcal{S}$ at which (10) is not satisfied. Does $\mathcal{S}$ admit a unique normal line at that point?
(d) Find the equation of the tangent plane $\mathcal{T}$ at $u = \frac{1}{2}$, $v = 0$. Sketch $\mathcal{T}$ and $\mathcal{S}$.
(e) Repeat part (d), for $u = \frac{1}{4}$, $v = 0$.
(f) Repeat part (d), for $u = \frac{1}{2}$, $v = \pi/4$.
(g) Repeat part (d), for $u = \frac{1}{2}$, $v = \pi/2$.

7. Consider the parametrization

$$x = a \cos u, \quad y = b \sin u, \quad z = v^2, \tag{7.1}$$

over $0 \le u < 2\pi$, $0 < v < 3$.

(a) Sketch $\mathcal{S}$, and verify that it is an elliptic cylinder.
(b) Show the $u = 0, \pi/4$, and $v = 2, 3$ coordinate curves on $\mathcal{S}$.

8. Give two different parametrizations of each surface:

(a) the x, y plane
(b) the x, z plane

(d) the plane $3x - 2y + z = 6$
(e) the hyperboloid $x^2 + y^2 - z^2 = 1$ over $0 \leq z < \infty$
(f) the elliptic paraboloid $x^2 + y^2 = z$
(g) the quadratic cone $x^2 + y^2 = z^2$

9. Find any parametrization $x = x(u,v)$, $y = y(u,v)$, $z = z(u,v)$ of the given plane such that the u, v coordinate curves are orthogonal.

(a) $x + 2y - 4z = 5$ (b) $2x - y + z = 9$
(c) $2x + y + 3z = -6$ (d) $x - y - z = 1$
(e) $x - y - 5z = 0$ (f) $x + y - 4z = 0$

10. Surely, if condition (10) is satisfied at some point P on a surface $\mathcal{S}$, then there exists a unique normal line and a unique tangent plane at P. Prove that the condition (10) is sufficient but not necessary. HINT: Logically, a counterexample will suffice. For such an example, consider the point $x = y = z = 0$ on the surface $z = 0$ (i.e., the x, y plane), with the parametrization $x = u^3$, $y = v^3$, $z = 0$.

11. Find a normal $\hat{\mathbf{n}}$ and the equation of the tangent plane for the given surface $\mathcal{S}$ and a point P; if a unique normal line and tangent plane do not exist at that point then state that.

(a) $\mathcal{S}: \ z = x^2 + y^2, \quad P = (2, 1, 5)$
(b) $\mathcal{S}: \ z = x^2 - 4y^2, \quad P = (1, 2, -3)$
(c) $\mathcal{S}: \ x^4 + y^4 + z^4 = 18, \quad P = (1, -1, 2)$
(d) $\mathcal{S}: \ x = yz, \quad P = (3, -1, -3)$
(e) $\mathcal{S}: \ y = x^2 + z^4, \quad P = (1, 2, 1)$
(f) $\mathcal{S}: \ x^2 + y^2 = 16\sin^2 z, \quad P = (2, 2, \pi/4)$
(g) $\mathcal{S}: \ x = u,\ y = u + v^2,\ z = v + 1, \quad P: \ u = 3,\ v = 2$
(h) $\mathcal{S}: \ x = u\cos v,\ y = u\sin v,\ z = u^2, \quad P: \ u = 2,\ v = \pi/6$
(i) $\mathcal{S}: \ x = u\sin v,\ y = u\cos v,\ z = u, \quad P: \ u = 0,\ v = \pi$
(j) $\mathcal{S}: \ x = ve^u,\ y = u - v,\ z = v^2, \quad P: \ u = 2,\ v = 1$
(k) $\mathcal{S}: \ x = u,\ y = u^2 + v,\ z = u - v, \quad P: \ u = 0,\ v = 1$
(l) $\mathcal{S}: \ x = 4u^2,\ y = u^2 - v^2,\ z = 2v^2 + 3, \quad P: \ u = 1,\ v = -1$

15.5 Surface Integrals

In Section 15.3.1 we reviewed double integrals $\iint f\, dA$ over regions in the x, y plane. Having developed the concept of the parametric representation of general curved surfaces in Section 15.4, we are now ready to generalize the concept of double integrals by considering double integrals on curved surfaces in 3-space. For example, suppose that we have a distribution of electric charges over a surface, and we wish to know the total charge or the electric field induced by that distribution. Such quantities are expressible as integrals of the sort that we are about to consider.

15.5.1. Area element dA. Consider a surface $\mathcal{S}$ given parametrically by $\mathbf{R}(u,v) = x(u,v)\hat{\mathbf{i}} + y(u,v)\hat{\mathbf{j}} + z(u,v)\hat{\mathbf{k}}$, where $\mathbf{R}(u,v)$ is C^1 (i.e., $\mathbf{R}$ and its first-order partial derivatives $\mathbf{R}_u$ and $\mathbf{R}_v$ are continuous) and $\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0}$ on $\mathcal{S}$. These conditions ensure that $\mathcal{S}$ is "nice." Specifically, they ensure that a unique normal line exists at each point on $\mathcal{S}$ and varies continuously on $\mathcal{S}$ (i.e., is a continuous function of u and v). Such a surface is said to be **smooth.** [illegible]

To obtain an expression for the area element dA on $\mathcal{S}$, in terms of our u, v coordinate system, we begin by taking a differential of the position vector $\mathbf{R}(u,v)$ from the origin to any point P on $\mathcal{S}$:

$$d\mathbf{R} = \mathbf{R}_u du + \mathbf{R}_v dv. \tag{1}$$

Thus, along the $u =$ constant and $v =$ constant curves through P we have $d\mathbf{R} = \mathbf{R}_v dv$ (since $u =$ constant, so that $du = 0$) and $d\mathbf{R} = \mathbf{R}_u du$, respectively, as shown in Fig. 1. These vectors define a parallelogram (Fig. 1) lying in the tangent plane to $\mathcal{S}$ at P. The area of this parallelogram is [recall (10) in Section 14.2]

$$dA = \|\mathbf{R}_u du \times \mathbf{R}_v dv\| = \|\mathbf{R}_u \times \mathbf{R}_v\|\, du\, dv. \tag{2}$$

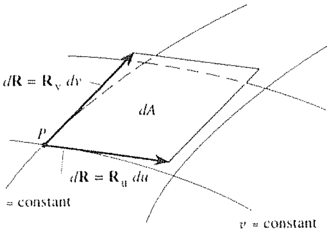

$d\mathbf{R} = \mathbf{R}_v\, dv$
dA
P
$d\mathbf{R} = \mathbf{R}_u\, du$
u = constant
v = constant

Figure 1. Area element dA.

As du and dv tend to zero this plane area element lies closer and closer to $\mathcal{S}$ so that it seems reasonable to define the **area** of the curved surface $\mathcal{S}$ by the double integral

$$\boxed{A = \iint_{\mathcal{S}} dA = \iint_{\mathcal{R}} \|\mathbf{R}_u \times \mathbf{R}_v\|\, du\, dv,} \tag{3}$$

where $\mathcal{R}$ is the region in the u, v plane that corresponds to the surface $\mathcal{S}$ in 3-space. Notice that we have not proved (3), it is a definition.* Further, we call

$$\boxed{dA = \|\mathbf{R}_u \times \mathbf{R}_v\|\, du\, dv} \tag{4}$$

the **area element** on $\mathcal{S}$.

To obtain a computational version of (4), cross $\mathbf{R}_u = x_u\hat{\mathbf{i}} + y_u\hat{\mathbf{j}} + z_u\hat{\mathbf{k}}$ with $\mathbf{R}_v = x_v\hat{\mathbf{i}} + y_v\hat{\mathbf{j}} + z_v\hat{\mathbf{k}}$. The norm of the resulting vector is the square root of the sum of the squares of its components. Carrying out these steps, we obtain

$$dA = \sqrt{EG - F^2}\, du\, dv, \tag{5a}$$

$$\begin{aligned} E &= x_u^2 + y_u^2 + z_u^2, \\ F &= x_u x_v + y_u y_v + z_u z_v, \\ G &= x_v^2 + y_v^2 + z_v^2. \end{aligned} \tag{5b}$$

Notice that we may attach a geometrical significance to F for, by inspection, we see that $F = \mathbf{R}_u \cdot \mathbf{R}_v$. Thus, if $\mathbf{R}(u, v)$ is such that F is identically zero, then that condition implies that $\mathbf{R}_u$ and $\mathbf{R}_v$ are perpendicular to each other at each point on $\mathcal{S}$. In that event, we say that the curvilinear coordinates u and v are **orthogonal**; that is, they form an "orthogonal net" on $\mathcal{S}$.

EXAMPLE 1. *Surface Area of a Sphere.* To illustrate the use of (5), let us compute the surface area of the spherical surface $\mathcal{S}$ defined by

$$x = a \sin v \cos u, \quad y = a \sin v \sin u, \quad z = a \cos v, \tag{6}$$

over $0 \le u \le \pi/2$ and $0 \le v \le \pi/2$ (which surface was the subject of Example 1 in Section 15.4). $\mathcal{S}$ constitutes one-eighth of a complete spherical surface and is shown in Fig. 2. Putting (6) into (5b) gives

$$E = a^2 \sin^2 v, \quad F = 0, \quad G = a^2. \tag{7}$$

*For instance, our use of the differential version of the chain rule, in (1), was heuristic, not rigorous.

Thus,

$$dA = a^2\,|\sin v|\,du\,dv = a^2 \sin v\,du\,dv \tag{8}$$

since $\sin v \geq 0$ over $0 \leq v \leq \pi/2$, so

$$A = a^2 \int_0^{\pi/2}\int_0^{\pi/2} \sin v\,du\,dv = \frac{\pi a^2}{2}. \tag{9}$$

Of course, (9) is correct since we know that the surface area of the sphere is $4\pi a^2$, and $\mathcal{S}$ is only one-eighth of that.

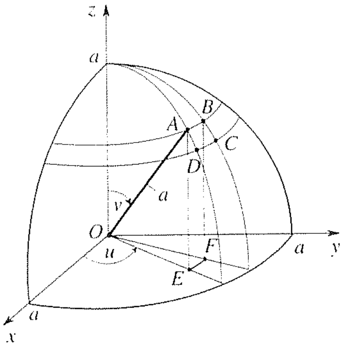


Figure 2. The surface $\mathcal{S}$.

COMMENT 1. The expression (8) could, in this case, have been obtained directly from Fig. 2, without recourse to (5). For dA is the shaded area $ABCD$ so

$$\begin{aligned} dA \sim (AB)(AD) &= (EF)(AD) = (OE\,du)(a\,dv) \\ &= (a \sin v\,du)(a\,dv) = a^2 \sin v\,du\,dv. \end{aligned} \tag{10}$$

Here AB, for example, denotes the length of the arc connecting the points A and B, du is the angle FOE, and dv is the angle DOA. But while the connection between (8) and the geometry contained in Fig. 2 is interesting and supportive, we should emphasize that the "method of pictures" used in (10) may not be feasible in other cases, whereas (5) is simple and automatic.

COMMENT 2. In this case $F = 0$ everywhere on $\mathcal{S}$ so the u, v coordinate curves must be *orthogonal*. This result is no great surprise since it should be evident that the curves of longitude and latitude intersect, at each point on $\mathcal{S}$, at right angles.

COMMENT 3. A striking feature of this example is the beautiful simplicity of the calculation in (9). This simplicity emphasizes that the u, v curvilinear coordinates defined in (6) are rather "natural" for the representation of this particular surface. ▮

There are two special cases of (5) that should be singled out:

Case 1: z=0. If $z = 0$, so that $\mathcal{S}$ is flat and lies in the x, y plane, then $x = x(u, v)$, $y = y(u, v)$, $z = 0$, and (5b) simplifies to

$$E = x_u^2 + y_u^2, \quad F = x_u x_v + y_u y_v, \quad G = x_v^2 + y_v^2. \tag{11}$$

It follows from (11) that

$$\begin{aligned} EG - F^2 &= x_u^2 y_v^2 - 2x_u y_u x_v y_v + x_v^2 y_u^2 \\ &= (x_u y_v - x_v y_u)^2, \end{aligned} \tag{12}$$

which is none other than the *Jacobian* $\partial(x, y)/\partial(u, v)$ squared! Thus, for the case where $z = 0$, (5) simplifies to the form

$$\boxed{dA = \left|\frac{\partial(x, y)}{\partial(u, v)}\right| du\,dv,} \tag{13}$$

the absolute value signs being included to ensure that we have the *positive* square root in (5a) since area must surely be positive.

EXAMPLE 2. As an illustration of Case 1, note that the parametrization choice

$$x = u\cos v, \qquad y = u\sin v, \qquad z = 0 \tag{14}$$

amounts to the familiar case of *plane polar coordinates*; that is, u is r and v is θ. Then the Jacobian is

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} = \begin{vmatrix} \cos v & -u\sin v \\ \sin v & u\cos v \end{vmatrix} = u,$$

so (13) gives

$$dA = |u|\, du\, dv$$

or, in terms of the more familiar letters r and θ,

$$\boxed{dA = r\, dr\, d\theta,} \tag{15}$$

where we omit the absolute values since the polar variable r is understood (in this book) to be nonnegative.

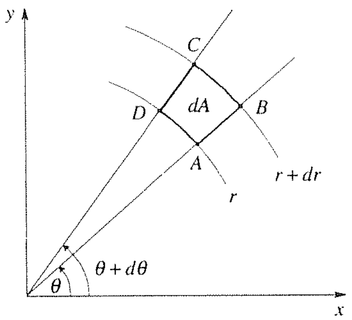


Figure 3. dA in polar coordinates.

Again, seeking a geometrical interpretation, note that the length of AB in Fig. 3 is dr, while the length of AD is $r\, d\theta$ so that $dA \sim (dr)(r\, d\theta) = r\, dr\, d\theta$, as in (15).

COMMENT. Since $dA = dx\, dy$ in Cartesian coordinates and $dA = r\, dr\, d\theta$ in polar coordinates, it is tempting to write $dx\, dy = r\, dr\, d\theta$. However, the latter is **not** correct. Heuristically, we can think of it this way. In the United States the smallest monetary unit is the penny so we can write d(money) = penny. In the country of Rumanova the smallest monetary unit is the yink so we can write d(money) = yink. Any amount of money can be formed by a suitable aggregation of pennies or of yinks. But it does not follow that a penny equals a yink. ▮

Case 2: $z = f(x, y)$. If $\mathcal{S}$ is given in the form

$$z = f(x, y), \tag{16}$$

over some region $\mathcal{R}$ in the x, y plane, then as noted in Section 15.4 (see Comment 2 in Example 1), we can parametrize $\mathcal{S}$ by

$$x = u, \quad y = v, \quad z = f(x, y). \tag{17}$$

Putting (17) into (5) yields $E = 1 + f_u^2$, $F = f_u f_v$, $G = 1 + f_v^2$, and hence the form

$$dA = \sqrt{1 + f_u^2 + f_v^2}\, du\, dv.$$

But since $x = u$ and $y = v$ it seems silly to retain the new variables u and v. Thus, let us write, instead,

$$\boxed{dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy.} \tag{18}$$

Correspondingly,

$$\boxed{A = \iint_{\mathcal{R}} \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy} \tag{19}$$

is the area of $\mathcal{S}$, where $\mathcal{R}$ is the region in the x, y plane lying directly beneath $\mathcal{S}$.

EXAMPLE 3. *Surface Area of a Sphere, Revisited.* Let us compute the surface area of the surface $\mathcal{S}$ shown in Fig. 2 (Example 1) again, this time using the parametrization

$$x = u, \qquad y = v, \qquad z = \sqrt{a^2 - u^2 - v^2}. \tag{20}$$

Since $x = u$ and $y = v$, the coordinate curves on $\mathcal{S}$ are simply the projections onto $\mathcal{S}$ of the $x =$ constant and $y =$ constant lines in the x, y plane. To illustrate, the coordinate curves through a given point P on $\mathcal{S}$ are sketched in Fig. 4. To visualize the coordinate curves more easily, imagine the $x =$ constant and $y =$ constant lines in the x, y plane (such as AB and CD, respectively) as thin wires, and imagine shining a flashlight upward, as shown in the figure. Then the shadow on $\mathcal{S}$ of the $x =$ constant lines will be the $u =$ constant curves (e.g., $A'B$), and the shadow on $\mathcal{S}$ of the $y =$ constant lines will be the $v =$ constant curves (e.g., $C'D$). Observe that the u, v curves are *not* orthogonal, in this case, since $F = f_u f_v$ is not zero, nor do they *look* orthogonal in Fig. 4.

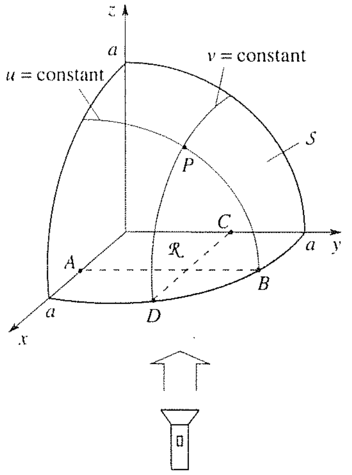


Figure 4. Visualizing the u, v coordinate curves on $\mathcal{S}$.

Returning to our calculation of the surface area of $\mathcal{S}$, we use (19), where $\mathcal{R}$ is the shaded quarter disk in the x, y plane (Fig. 4). Since $f_x = (\frac{1}{2})(-2x)(a^2 - x^2 - y^2)^{-1/2}$ and $f_y = (\frac{1}{2})(-2y)(a^2 - x^2 - y^2)^{-1/2}$, (19) gives

$$A = \int_0^a \int_0^{\sqrt{a^2 - y^2}} \frac{a}{\sqrt{a^2 - x^2 - y^2}}\, dx\, dy. \tag{21}$$

Evaluation of the iterated integral in (21) is left for the exercises; the answer, of course, is $\pi a^2/2$, as before. ▮

15.5.2. Surface integrals. We are ready to study **surface integrals**, integrals of the form

$$\iint_{\mathcal{S}} f\, dA, \tag{22}$$

where the function f is known on the surface $\mathcal{S}$.

Our definition of the surface integral (22) will be essentially the same as that given in Section 15.3.1 for the double integral $\iint_{\mathcal{R}} f\, dA$ on some region $\mathcal{R}$ in the x, y plane. That is, we partition $\mathcal{S}$ into N parts $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_N$, and define the *norm* of the partition $|p|$ as the "size" of the largest one of these parts. (As a measure of the "size" of an element $\mathcal{S}_j$, we can, for example, use the least upper bound of the linear distances between all possible pairs of points on $\mathcal{S}_j$.) Next, let (x_j, y_j, z_j) be an arbitrarily selected point on $\mathcal{S}_j$ for each $j = 1, 2, \dots, N$. Then, denoting the area of $\mathcal{S}_j$ as ΔA_j,

$$\sum_{j=1}^{N} f(x_j, y_j, z_j)\Delta A_j \tag{23}$$

is called the **Riemann sum** corresponding to the chosen partition and (x_j, y_j, z_j) points. The idea, in principle at least, is to compute the Riemann sum (23), then to introduce a finer partition (i.e., one with a smaller norm) and compute the new Riemann sum, and so on, such that the norm of the partitions tends to zero. If the sequence of values of the Riemann sum thus generated converges to a unique limit, independent of the choice of the partition sequence and (x_j, y_j, z_j) points, that limit is, by definition, the **surface integral** $\iint_{\mathcal{S}} f\, dA$; that is,

$$\iint_{\mathcal{S}} f(x,y,z)\, dA \equiv \lim_{|p|\to 0} \sum f(x_j, y_j, z_j)\Delta A_j. \tag{24}$$

To *evaluate* a surface integral we do *not* use the definition (24) (although we *could* carry out that exercise in sufficiently simple cases). Rather, we convert the integral to an iterated integral by parametrizing $\mathcal{S}$ in the form $\mathbf{R}(u,v)$ and using the expression $dA = \|\mathbf{R}_u \times \mathbf{R}_v\|\, du\, dv = \sqrt{EG - F^2}\, du\, dv$. Thus,

$$\iint_{\mathcal{S}} f\, dA = \iint_{\mathcal{R}} f(x(u,v), y(u,v), z(u,v))\sqrt{EG - F^2}\, du\, dv, \tag{25a}$$

where $\mathcal{R}$ is the region in the u, v plane that corresponds to the surface $\mathcal{S}$ in 3-space. Or if $\mathcal{S}$ is given in the form $z = z(x,y)$, then

$$\iint_{\mathcal{S}} f\, dA = \iint_{\mathcal{R}} f(x, y, z(x,y))\sqrt{1 + z_x^2 + z_y^2}\, dx\, dy, \tag{25b}$$

where $\mathcal{R}$ is the "shadow" of $\mathcal{S}$ on the x, y plane.

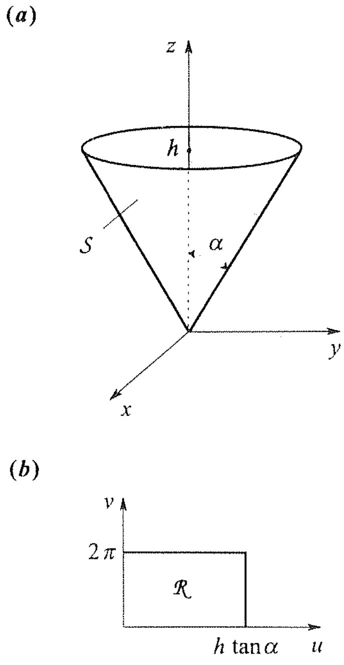


Figure 5. Conical shell.

EXAMPLE 4. *Mass of a Conical Shell.* Consider a thin shell in the shape of a right circular cone $\mathcal{S}$ of height h as shown in Fig. 5a. Assuming that the mass density distribution σ (mass per unit surface area) is known, we wish to compute the total mass, which may be expressed as the surface integral

$$M = \iint_{\mathcal{S}} \sigma\, dA. \tag{26}$$

First, let us parametrize $\mathcal{S}$. (Before reading on, you should try to make up a suitable parametrization, remembering that the parametrization of a given surface is not unique, and that the rule of thumb is to construct a parametrization which is as simple and natural as possible.) Adopting the parametrization

$$\begin{aligned} x &= u\cos v, \\ y &= u\sin v, \\ z &= u\cot\alpha, \end{aligned} \tag{27}$$

over the region $\mathcal{R}$ shown in Fig. 5b, we compute $E = x_u^2 + y_u^2 + z_u^2 = 1 + \cot^2\alpha = 1/\sin^2\alpha$, $F = x_u x_v + y_u y_v + z_u z_v = 0$, and $G = x_v^2 + y_v^2 + z_v^2 = u^2$, so that the form (25b) gives

$$M = \iint\limits_{\mathcal{R}} \sigma(u,v)\,\frac{u}{\sin\alpha}\,du\,dv = \frac{1}{\sin\alpha}\int_0^{2\pi}\int_0^{h\tan\alpha} \sigma(u,v)u\,du\,dv. \tag{28}$$

For example, if $\sigma(u,v) = \text{constant} = \sigma_0$, in which case we say that the shell is homogeneous, then

$$M = \frac{\sigma_0}{\sin\alpha}\int_0^{2\pi}\int_0^{h\tan\alpha} u\,du\,dv = \pi\sigma_0 h^2 \sin\alpha\sec^2\alpha. \tag{29}$$

COMMENT 1. Where did we get (27)? In cylindrical coordinates the equation of the cone is $z = r\cot\alpha$. Also, $x = r\cos\theta$ and $y = r\sin\theta$. These three equations are a suitable parametrization of $\mathcal{S}$, where the parameters are r and θ. Finally, we took r to be u and θ to be v, to be consistent with our usual u, v notation. If we chose $r = v$ and $\theta = u$, instead, the same final result (29) would be obtained.

COMMENT 2. Alternatively, spherical polars are an attractive choice because the cone is a constant ϕ surface, namely, $\phi = \alpha$. Thus, $x = \rho\sin\phi\cos\theta = \rho\sin\alpha\cos\theta$, $y = \rho\sin\phi\sin\theta = \rho\sin\alpha\sin\theta$, and $z = \rho\cos\phi = \rho\cos\alpha$ so if we let ρ be u and θ be v, say, then we have the alternative parametrization

$$\begin{aligned} x &= u\sin\alpha\cos v, \\ y &= u\sin\alpha\sin v, \\ z &= u\cos\alpha. \end{aligned} \tag{30}$$

From (30) we obtain $E = 1$, $F = 0$, and $G = u^2\sin^2\alpha$ so (with $\sigma = \sigma_0$ again)

$$M = \int_0^{2\pi}\int_0^{h\sec\alpha} \sigma_0 u \sin\alpha\,du\,dv, \tag{31}$$

which gives the same final result as obtained in (29).

COMMENT 3. As a partial check of (29), observe that $M \to 0$ as $\alpha \to 0$, as it should, and $M \to \infty$ as $\alpha \to \pi/2$, as it should. ▮

Closure. The key to this section is the definition, by (4), of the area element dA of a smooth surface $\mathcal{S}$ that is defined parametrically in terms of u and v. In turn, (4) springs from Fig. 1, which should be understood and remembered. Working out $\|\mathbf{R}_u \times \mathbf{R}_v\|$ in terms of $x(u,v)$, $y(u,v)$, $z(u,v)$ then produce the computationally convenient version (5). In addition, we call attention to two important special cases. In the first, $\mathcal{S}$ is flat and lies in the x, y plane and dA reduces to the absolute magnitude of the Jacobian times $du\,dv$; in the second, the shape of $\mathcal{S}$ is known in the form $z = f(x,y)$, and dA then simplifies to the form given in (18).

EXERCISES 15.5

1. In each case evaluate the surface area A of the surface $\mathcal{S}$ using (5).

(a) Let $\mathcal{S}$ be the *circular cylinder* $x^2 + y^2 = 1$, between $z = 0$ and $z = 1 + y$. HINT: Use $x = \cos v$, $y = \sin v$, $z = u$.
(b) Let $\mathcal{S}$ be the *circular cylinder* $x^2 + y^2 = 1$, between $z = 0$ and $z = 1 - y^2$. HINT: Same hint as in part (a).
(c) Let $\mathcal{S}$ be the *quadric cone* $x^2 + y^2 = z^2$, between $z = 0$ and $z = h$. Show that $A = \sqrt{2}\pi h^2$. HINT: Use $x = u \cos v$, $y = u \sin v$, $z = u$.

2. Let $\mathcal{S}$ be the *elliptic paraboloid* $x^2 + y^2 = z$, between $z = 0$ and $z = h$.

(a) Using (5), show that $A = \left(\frac{\pi}{6}\right)[(1 + 4h)^{3/2} - 1]$.
(b) Applying Taylor's series to the answer in part (a), show that $A \sim \pi h$ as $h \to 0$. Explain why this result looks correct [and hence provides us with a partial check on our answer to part (a)].

3. Let $\mathcal{S}$ be the one-sheeted *hyperboloid* $x^2 + y^2 - z^2 = 1$, between $z = 0$ and $z = h$.

(a) Show that

$$A = \pi\left[h\sqrt{1 + 2h^2} + \frac{1}{\sqrt{2}}\ln\left|\sqrt{2}\,h + \sqrt{1 + 2h^2}\right|\right]. \tag{3.1}$$

(b) Applying Taylor's series to (3.1), show that $A \sim 2\pi h$ as $h \to 0$. Explain why this result looks correct [and hence provides us with a partial check on (a)].

4. Let $\mathcal{S}$ be the *hyperboloid* $z^2 - x^2 - y^2 = 1$, between $z = 1$ and $z = h$.

(a) Using (5), with $x = u \cos v$, $y = u \sin v$, $z = \sqrt{u^2 + 1}$, show that

$$A = \pi\Big[h\sqrt{2h^2 - 1} - \tfrac{1}{\sqrt{2}}\ln\left|\sqrt{2}\,h + \sqrt{2h^2 - 1}\right| \\ -1 + \tfrac{1}{\sqrt{2}}\ln\left|1 + \sqrt{2}\right|\Big]. \tag{4.1}$$

HINT: Same hint as in Exercise 3(a).

(b) To examine the case where $\mathcal{S}$ is very "shallow," set $h = 1 + \epsilon$ in (4.1), and consider the behavior of A as $\epsilon \to 0$. Explain why the result looks correct [and hence provides us with a partial check on (4.1)].

(c) In the same spirit as part (b), show that $A \sim \sqrt{2}\pi h^2$ as $h \to \infty$. Explain why this result looks correct [and therefore provides us with another partial check on (4.1)].

5. Parametrizing the circular cylinder $x^2 + y^2 = a^2$ by $x = a \cos v$, $y = a \sin v$, $z = u$, show that the area element is $dA = a\,du\,dv$. Interpret this result geometrically, with a labeled sketch.

6. Find the area A of the following regions in the x, y plane:

(a) the region enclosed by $r = \sin\theta$ $(0 \le \theta < \pi)$
(b) the region enclosed by the limaçon $r = 2 + \cos\theta$ $(0 \le \theta < 2\pi)$
(c) the region enclosed by one leaf of the "daisy" $r = 2\sin 3\theta$
(d) the region (in the second quadrant) between the circle $r = \sin\theta$ $(0 \le \theta < \pi)$ and the cardioid $r = 1 + \cos\theta$ $(0 \le \theta < 2\pi)$.

7. Consider a portion $\mathcal{S}$ of a plane $ax + by + cz = d$ with a "shadow" on the x, y plane designated as the region $\mathcal{R}$. We assume that the plane is not perpendicular to the x, y plane, so that $c \ne 0$ and $z = (d - ax - by)/c$. Using (19), show that the area of $\mathcal{S}$ is

$$A = (\sec\alpha)(\text{area of } \mathcal{R}), \tag{7.1}$$

where α is the acute angle between the z axis and the normal line to $\mathcal{S}$.

8. Use (19) to show that the surface area of the paraboloid $z = h(1 - x^2 - y^2)$, between $z = 0$ and $z = h$, is

$$A = \iint_{\mathcal{R}} \sqrt{1 + 4h^2(x^2 + y^2)}\,dx\,dy, \tag{8.1}$$

where $\mathcal{R}$ is the unit disk $x^2 + y^2 \le 1$. To integrate (8.1), change to polar coordinates r, θ, and show that

$$A = \frac{\pi}{6h^2}\left[(1 + 4h^2)^{3/2} - 1\right]. \tag{8.2}$$

As a partial check on (8.2), show that the right-hand side of (8.2) tends to π (namely, the area of the unit disk) as $h \to 0$.

9. Sketch the surface $\mathcal{S}$ defined by $z = \epsilon(1 - x^2)(1 - y^2)$ over the square $0 \le x \le 1$, $0 \le y \le 1$, where $0 \le \epsilon \ll 1$. Show that (19) gives the area of $\mathcal{S}$ as

$$A = \int_0^1\int_0^1 \sqrt{1 + 4\epsilon^2\left[x^2(1 - y^2)^2 + (1 - x^2)^2y^2\right]}\,dx\,dy.$$

(9.1)

Evidently, this integral is difficult to evaluate. However, recalling that ϵ is small, suppose that we set $t \equiv 4\epsilon^2\left[x^2(1-y^2)^2+(1-x^2)^2y^2\right]$ and expand the integrand as a Maclaurin series in t:

$$\begin{aligned}\sqrt{1+4\epsilon^2\left[x^2(1-y^2)^2+(1-x^2)^2y^2\right]}&\\ &=(1+t)^{1/2}\\ &=1+\frac{1}{2}t-\frac{1}{8}t^2+\cdots \qquad (|t|<1)\\ &=1+2\epsilon^2[x^2(1-y^2)^2+(1-x^2)^2y^2]\\ &\quad -2\epsilon^4[x^2(1-y^2)^2+(1-x^2)^2y^2]^2+\cdots.\end{aligned} \tag{9.2}$$

Replacing the integrand in (9.1) by the infinite series on the right-hand side of (9.2), we succeed in trading in one function that is very difficult to integrate for many that are quite simple to integrate (since they are of the form x^my^n). In fact, we do not even need many terms, for good accuracy, if ϵ is sufficiently small, for then $t \ll 1$ so that $(1+t)^{1/2} \approx 1+\frac{1}{2}t$ or $1+\frac{1}{2}t-\frac{1}{8}t^2$, say, may suffice. Using $(1+t)^{1/2}\approx 1+\frac{1}{2}t$, show that $A \approx 1+\frac{32}{45}\epsilon^2$. NOTE: The preceding discussion does not pretend to be any more than heuristic. A rigorous basis for integrating the infinite series in a term-by-term manner is provided by the following *theorem*: If $|a_n(x,y)| < M_n$ over the closed region $\mathcal{R}$, where the M_n's are constants and $\sum^\infty M_n$ is convergent, then

$$\iint_{\mathcal{R}} \sum^\infty a_n(x,y)\,dx\,dy = \sum^\infty \iint_{\mathcal{R}} a_n(x,y)\,dx\,dy \tag{9.3}$$

(i.e., we may integrate the series term by term).

10. Evaluate $\iint_{\mathcal{S}}(1+x)dA$, where the surface $\mathcal{S}$ is

(a) the plane $z = 1+y$ with vertices at $(0,0,1)$, $(1,0,1)$, $(1,1,2)$, and $(0,1,2)$
(b) the plane $z = x+y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,1)$
(c) the cylinder $x^2+y^2=1$, between $z=0$ and $z=h$
(d) the cylinder $x^2+y^2=1$, between $z=0$ and $z=1+x$
(e) the hemisphere $x^2+y^2+z^2=9$, between $x=0$ and $x=3$
(f) the sphere $x^2+y^2+z^2=4$
(g) the quadric cone $x^2+y^2=z^2$, between $z=0$ and $z=h$
HINT: Use $x=u\cos v$, $y=u\sin v$, $z=u$.
(h) elliptic paraboloid $x^2+y^2=z$, between $z=0$ and $z=h$

11. (*Mass and center of gravity*) Let σ be the mass density of a (negligibly thick) distribution of mass over a surface $\mathcal{S}$. That is, σ is the mass per unit area at each point on $\mathcal{S}$; it may vary over $\mathcal{S}$. Then the x, y, z coordinates of the **center of gravity** are defined as

$$\begin{aligned}x_c &= \frac{1}{M}\iint_{\mathcal{S}} x\sigma\,dA,\\ y_c &= \frac{1}{M}\iint_{\mathcal{S}} y\sigma\,dA,\\ z_c &= \frac{1}{M}\iint_{\mathcal{S}} z\sigma\,dA,\end{aligned} \tag{11.1}$$

where

$$M = \iint_{\mathcal{S}} \sigma\,dA \tag{11.2}$$

is the total mass. Evaluate x_c in each case. (You need not evaluate y_c, z_c.)

(a) $\mathcal{S}$ is the plane surface $z = x+2y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,2)$; $\sigma = \text{constant} = \sigma_0$.
(b) $\mathcal{S}$ is the plane surface $z = x+y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,1)$; $\sigma = 1+y$.
(c) $\mathcal{S}$ is the plane surface $z = 2-x$ with vertices at $(0,0,2)$, $(0,1,2)$, $(2,0,0)$, $(2,1,0)$; $\sigma = 4-x$.
(d) $\mathcal{S}$ is the plane surface $z = 2-x$ with vertices at $(0,0,2)$, $(0,1,2)$, $(2,0,0)$; $\sigma = 1+x$.
(e) $\mathcal{S}$ is the cylindrical surface $x^2+y^2=4$ between $z=0$ and $z=h$; $\sigma = 4+x$.
(f) $\mathcal{S}$ is the cylindrical surface $x^2+y^2=4$ between $z=0$ and $z=2+x$; $\sigma = \text{constant} = \sigma_0$.
(g) $\mathcal{S}$ is the hemispherical surface $x^2+y^2+z^2=1$ between $x=0$ and $x=1$; $\sigma = \text{constant} = \sigma_0$.
(h) $\mathcal{S}$ is the spherical surface $x^2+y^2+z^2=9$; $\sigma = 4+x$.

12. (*Moments of inertia*) Let σ be the mass density of a (negligibly thick) distribution of mass over a surface $\mathcal{S}$, as in Exercise 11. Then the **moments of inertia** I_x, I_y, I_z about the x, y, z axes, respectively, are defined as

$$\begin{aligned}I_x &= \iint_{\mathcal{S}} (y^2+z^2)\sigma\,dA,\\ I_y &= \iint_{\mathcal{S}} (x^2+z^2)\sigma\,dA,\\ I_z &= \iint_{\mathcal{S}} (x^2+y^2)\sigma\,dA.\end{aligned} \tag{12.1}$$

Evaluate I_x in each case:

(a) same as in Exercise 11(a)

(b) same as in Exercise 11(b)
(c) same as in Exercise 11(c)
(d) S is the spherical surface $x^2+y^2+z^2=c^2$; $\sigma=$ constant $=\sigma_0$
(e) S is the cylindrical surface $y^2+z^2=4$ between $x=0$ and $x=3$; $\sigma=$ constant $=\sigma_0$
(f) S is a flat disk of radius c, lying in the y,z plane, with its center at the origin; $\sigma=$ constant $=\sigma_0$
(g) S is a flat disk of radius c, lying in the x,z plane, with its center at the origin; $\sigma=$ constant $=\sigma_0$.

13. In each case S is flat and lies in the x,y plane. Evaluate each integral by introducing polar coordinates r,θ.

(a) $\int\int_S \frac{y}{x}\,dx\,dy,\quad S\colon 1\le r\le 2,\ 0\le\theta\le\pi/4$
(b) $\int\int_S \sin(x^2+y^2)\,dx\,dy,\quad S\colon r\le 2,\ 0\le\theta\le\pi/3$
(c) $\int\int_S \frac{y(x^2+y^2)^2}{x}\,dx\,dy,\quad S\colon 1\le r\le 2,$ $-\pi/3\le\theta\le\pi/3$
(d) $\int\int_S (x^2+y^2)^{-1/2}\,dx\,dy,\quad S\colon 1\le r\le 2,\ 0\le\theta\le\pi/2$
(e) $\int\int_S x^2(x^2+y^2)^{-3/2}\,dx\,dy,\quad S$: triangle with vertices at $(0,0)$, $(1,0)$, and $(1,1)$
(f) $\int\int_S e^{-(x^2+y^2)}\,dx\,dy,\quad S\colon r\le 1,\ 0\le\theta\le\pi$

14. To evaluate

$$I=\int_0^1\int_0^{1-y} e^{x/(x+2y)}\,dx\,dy, \tag{14.1}$$

make the change of variables

$$u=x, \qquad v=x+2y. \tag{14.2}$$

The resulting integrand will be an easy function of u and a hard function of v so integrate on u first. To determine the u,v integration limits first infer the region of integration in the x,y plane, from (14.1), then use the transformation (14.2) to determine the region of integration in the u,v plane. Thus, show that

$$I=\frac{e}{4}-1+\frac{1}{2e}\int_1^2 ve^{2/v}\,dv. \tag{14.3}$$

Finally, use computer software to evaluate the v integral in (14.3), and thus show that $I=0.76624517$. HINT: For numerical integration using *Maple*, see ?int[numerical].

15. (a) To evaluate

$$I=\int_0^1\int_0^1 \sin\frac{xy}{xy+1}\,dx\,dy, \tag{15.1}$$

make the change of variables $u=x$, $v=xy$ and, using the ideas outlined in Exercise 14, show that $I=0.174971978$.
(b) In part (a), the choice $u=xy$, $v=xy+1$ might have seemed more natural. Show why that choice does *not* work.

15.6 Volumes and Volume Integrals

Having discussed the generation of curves by one-parameter families of the form $x=x(u)$, $y=y(u)$, $z=z(u)$, and the generation of surfaces by two-parameter families of the form $x=x(u,v)$, $y=y(u,v)$, $z=z(u,v)$, it should be no surprise that we now consider the generation of *volumes* by three-parameter families of the form

$$x=x(u,v,w), \qquad y=y(u,v,w), \qquad z=z(u,v,w), \tag{1}$$

or in terms of the position vector $\mathbf{R}$,

$$\mathbf{R}(u,v,w)=x(u,v,w)\hat{\mathbf{i}}+y(u,v,w)\hat{\mathbf{j}}+z(u,v,w)\hat{\mathbf{k}}. \tag{2}$$

That is, for each fixed w the parametrization (1) [or (2)] defines a surface (in general). Thus, as we vary w we produce a family of such surfaces which, in general, will generate a volume.

15.6.1. Volume element dV. We assume that $\mathbf{R}(u,v,w)$ is C^1 in the region $\mathcal{V}$ of interest (i.e., $\mathbf{R}$, $\mathbf{R}_u$, $\mathbf{R}_v$, and $\mathbf{R}_w$ are continuous in $\mathcal{V}$), and that $\mathbf{R}_u$, $\mathbf{R}_v$, $\mathbf{R}_w$ are linearly independent (or, equivalently, that $\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w \neq 0$) at each point in $\mathcal{V}$. Mimicing our discussion of surface area in Section 15.5.1, consider the $u =$ constant, $v =$ constant, and $w =$ constant surfaces through a given point P as sketched in Fig. 1. If the curve PQ is the intersection of the $w =$ constant, and $v =$ constant surfaces through P, then only u varies along PQ so PQ is the u coordinate curve through P. Similarly, PR is the v coordinate curve through P, and PS is the w coordinate curve through P.

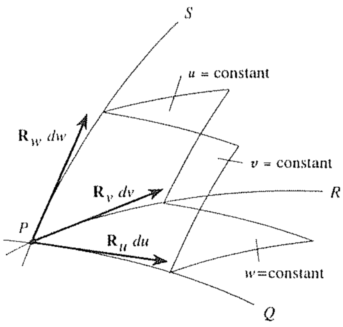


Figure 1. Edges of the parallelepiped for dV.

Since $\mathbf{R}_u$, $\mathbf{R}_v$, $\mathbf{R}_w$ are linearly independent, by assumption, the vectors $\mathbf{R}_u\,du$, $\mathbf{R}_v\,dv$, $\mathbf{R}_w\,dw$ determine a parallelepiped of nonzero volume dV. According to (5) in Section 14.4,

$$\begin{aligned} dV &= |\mathbf{R}_u\,du \cdot \mathbf{R}_v\,dv \times \mathbf{R}_w\,dw| = |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\,du\,dv\,dw \\ &= \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\,dv\,dw, \end{aligned} \tag{3}$$

where the inner vertical rules on the right-hand side of (3) denote determinant, and the outer ones denote absolute value. But the determinant in (3) is none other than the Jacobian $\partial(x,y,z)/\partial(u,v,w)$* so that

$$\boxed{dV = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw,} \tag{4}$$

and this quantity is hereby *defined* to be the **volume element** at P. Then the volume V of a given region $\mathcal{V}$ in x, y, z space may be expressed as

$$\boxed{V = \iiint_{\mathcal{V}} dV = \iiint_{\mathcal{R}} \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw,} \tag{5}$$

where $\mathcal{R}$ is the region in u, v, w space corresponding to the region $\mathcal{V}$ in x, y, z space.

EXAMPLE 1. *Cylindrical Coordinates.* If we let the cylindrical coordinates r, θ, z be u, v, w, respectively, then

$$x = u\cos v, \qquad y = u\sin v, \qquad z = w \tag{6}$$

*Recall that $\det \mathbf{A}^{\mathrm{T}} = \det \mathbf{A}$ for any square matrix $\mathbf{A}$. Therefore, the Jacobian $J(u,v,w) = \dfrac{\partial(x,y,z)}{\partial(u,v,w)}$ can be expressed either as

$$\begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} \quad \text{or as} \quad \begin{vmatrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{vmatrix}.$$

is our parametrization. Then

$$\frac{\partial(x,y,z)}{\partial(u,v,w)} = \begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} = \begin{vmatrix} \cos v & -u\sin v & 0 \\ \sin v & u\cos v & 0 \\ 0 & 0 & 1 \end{vmatrix} = u \tag{7}$$

so (4) gives $dV = |u|\, du\, dv\, dw$ or, returning to the more familiar r, θ, z notation, $dV = |r|\, dr\, d\theta\, dz$. Since $r \geq 0$, we have the result

$$\boxed{dV = r\, dr\, d\theta\, dz,} \tag{8}$$

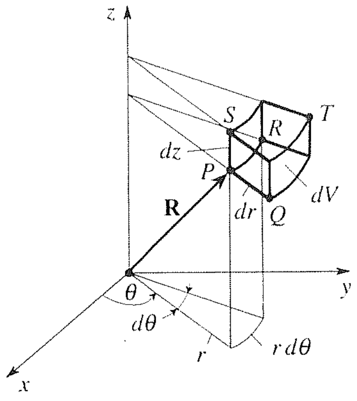


Figure 2. Cylindrical coordinates.

which result admits the simple geometrical interpretation indicated in Fig. 2. To explain, let us write down the position vector [recall (22) in Section 14.6]

$$\mathbf{R} = r\hat{\mathbf{e}}_r(\theta) + z\hat{\mathbf{e}}_z. \tag{9}$$

Then

$$\begin{aligned} d\mathbf{R} &= dr\hat{\mathbf{e}}_r + r\frac{d\hat{\mathbf{e}}_r}{d\theta}\, d\theta + dz\hat{\mathbf{e}}_z \\ &= dr\hat{\mathbf{e}}_r + r\, d\theta\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z \end{aligned} \tag{10}$$

or, in the notation of Fig. 2,

$$\mathbf{PT} = \mathbf{PQ} + \mathbf{PR} + \mathbf{PS}. \tag{11}$$

Since $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ are orthogonal, it follows that $dV = \|\mathbf{PQ}\|\, \|\mathbf{PR}\|\, \|\mathbf{PS}\| = (dr)(r\, d\theta)(dz)$, as given by (8). ▮

EXAMPLE 2. *Spherical Coordinates.* If we let the spherical coordinates ρ, ϕ, θ be u, v, w, respectively, then

$$\begin{aligned} x &= u\sin v\cos w, \\ y &= u\sin v\sin w, \\ z &= u\cos v, \end{aligned} \tag{12}$$

so

$$\begin{aligned} \frac{\partial(x,y,z)}{\partial(u,v,w)} &= \begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} \\ &= \begin{vmatrix} \sin v\cos w & u\cos v\cos w & -u\sin v\sin w \\ \sin v\sin w & u\cos v\sin w & u\sin v\cos w \\ \cos v & -u\sin v & 0 \end{vmatrix} \\ &= u^2\sin v. \end{aligned} \tag{13}$$

Thus, (4) gives $dV = \left|u^2 \sin v\right| du\, dv\, dw = \rho^2 \left|\sin\phi\right| d\rho\, d\phi\, d\theta$. As noted in Section 14.6.3, we normally limit ϕ and θ so that $0 \le \phi \le \pi$ and $0 \le \theta < 2\pi$. Since $0 \le \sin\phi \le 1$ on $0 \le \phi \le \pi$ we can drop the absolute value signs and write

$$\boxed{dV = \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta.} \tag{14}$$

Alternatively,

$$\mathbf{R} = \rho \hat{\mathbf{e}}_\rho(\phi, \theta) \tag{15}$$

so

$$\begin{aligned} d\mathbf{R} &= d\rho \hat{\mathbf{e}}_\rho + \rho\left(\frac{\partial \hat{\mathbf{e}}_\rho}{\partial \phi} d\phi + \frac{\partial \hat{\mathbf{e}}_\rho}{\partial \theta} d\theta\right) \\ &= d\rho \hat{\mathbf{e}}_\rho + \rho d\phi \hat{\mathbf{e}}_\phi + \rho \sin\phi\, d\theta \hat{\mathbf{e}}_\theta, \end{aligned} \tag{16}$$

where we have used (28) in Section 14.6. And since $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ are orthogonal, it follows from (16) that $dV = (d\rho)(\rho\, d\phi)(\rho \sin\phi\, d\theta)$ as given by (14). For geometrical interpretation see Fig. 3. ▮

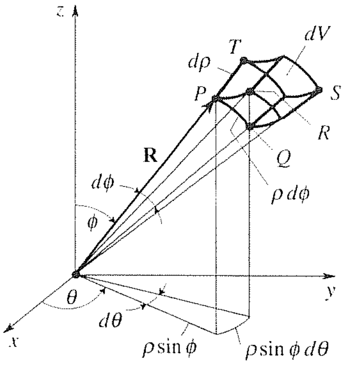


Figure 3. Spherical coordinates.

To tie the present section together with the material in Section 14.6 on cylindrical and spherical coordinates, observe that for those coordinate systems we can derive the *volume element* not only from (4), but also from the product of the three orthogonal components of the $d\mathbf{R}$ vector. Furthermore, the products of those components taken two at a time give the *area elements* on the constant-coordinate surfaces.

For instance, on a conical $\phi =$ constant surface $d\phi = 0$ so (16) becomes $d\mathbf{R} = d\rho \hat{\mathbf{e}}_\rho + \rho \sin\phi\, d\theta\, \hat{\mathbf{e}}_\theta$, and $dA = (d\rho)(\rho \sin\phi\, d\theta) = \rho \sin\phi\, d\rho\, d\theta$ is the relevant area element, namely, the shaded area in Fig. 3. Similarly, on a spherical $\rho =$ constant surface $d\rho = 0$ so (16) becomes $d\mathbf{R} = \rho\, d\phi \hat{\mathbf{e}}_\phi + \rho \sin\phi\, d\theta\, \hat{\mathbf{e}}_\theta$, and $dA = (\rho\, d\phi)(\rho \sin\phi\, d\theta) = \rho^2 \sin\phi\, d\phi\, d\theta$. Of course, these results are the same as would be obtained using the $dA = \sqrt{EG - F^2}\, du\, dv$ formula in Section 15.5.

Thus, since it contains so much information, the expression for $d\mathbf{R}$ for any given orthogonal coordinate system, is extremely important. Let us summarize these results, for reference, for cylindrical and spherical coordinates.

Cylindrical coordinates:

$$\boxed{\begin{aligned} \mathbf{R} &= r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z \\ d\mathbf{R} &= dr\hat{\mathbf{e}}_r + r d\theta \hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z \\ dA &= \begin{cases} r\, d\theta\, dz & \text{(constant-}r\text{ surface)} \\ dr\, dz & \text{(constant-}\theta\text{ surface)} \\ r\, dr\, d\theta & \text{(constant-}z\text{ surface)} \end{cases} \\ dV &= r\, dr\, d\theta\, dz. \end{aligned}} \tag{17}$$

Spherical coordinates:

$$\boxed{\begin{aligned} \mathbf{R} &= \rho\hat{\mathbf{e}}_\rho \\ d\mathbf{R} &= d\rho\hat{\mathbf{e}}_\rho + \rho\, d\phi\hat{\mathbf{e}}_\phi + \rho\sin\phi\, d\theta\hat{\mathbf{e}}_\theta \\ dA &= \begin{cases} \rho^2\,|\sin\phi|\,d\phi\,d\theta & \text{(constant-}\rho\text{ surface)} \\ \rho\,|\sin\phi|\,d\rho\,d\theta & \text{(constant-}\phi\text{ surface)} \\ \rho d\rho\,d\phi & \text{(constant-}\theta\text{ surface)} \end{cases} \\ dV &= \rho^2\,|\sin\phi|\,d\rho\,d\phi\,d\theta, \end{aligned}} \tag{18}$$

where the absolute value signs are to be enforce that the area and volume elements always be positive.

15.6.2. Volume integrals. With dV in hand, as given by (4), we can deal with the **volume integral** $\iiint_{\mathcal{V}} f\,dV$ of a given function f over a given region $\mathcal{V}$ in 3-space for if x, y, z are parametrized by u, v, w, then

$$\boxed{\int_{\mathcal{V}} f\,dV = \int_{\mathcal{R}} f(x(u,v,w),\ y(u,v,w),\ z(u,v,w)) \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\,dv\,dw,} \tag{19}$$

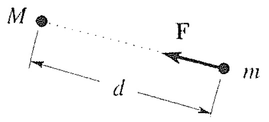


Figure 4. Gravitational force $\mathbf{F}$ exerted by M on m.

where $\mathcal{R}$ is the region in u, v, w space corresponding to the region $\mathcal{V}$ in x, y, z space. We use single integral signs in (19) to denote triple integrals, for compactness, and use that notation routinely in Chapter 16. There should be no confusion, in (19), because of the $\mathcal{V}$ and $\mathcal{R}$, which are three-dimensional regions, and because the dV and $du\,dv\,dv$ clearly denote triple integrals.

EXAMPLE 3. *Gravitational Attraction of a Cone.* **Newton's law of gravitation** states that the force of attraction $\mathbf{F}$ exerted by any one point mass M on any other point mass m is given by

$$\mathbf{F} = G\frac{Mm}{d^2}\hat{\mathbf{e}}, \tag{20}$$

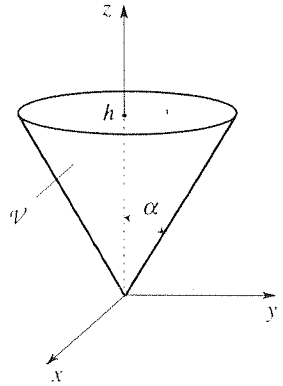


Figure 5. The cone.

where (Fig. 4) d is the distance of separation, $\hat{\mathbf{e}}$ is a unit vector directed from m toward M, and G ($= 6.67 \times 10^{-8}$ cm^3/g sec^2) is the universal gravitational constant; (20) is said to be an *inverse square law* since the force varies as the inverse square of the distance. (By M and m being *point masses*, we mean that their sizes are negligible compared with d.) If M is at (x, y, z) and m is at (x_0, y_0, z_0), we can express $\hat{\mathbf{e}}$ as

$$\hat{\mathbf{e}} = \frac{(x-x_0)\hat{\mathbf{i}} + (y-y_0)\hat{\mathbf{j}} + (z-z_0)\hat{\mathbf{k}}}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}}. \tag{21}$$

The problem that we pose here is to calculate the force of attraction exerted at the origin, per unit mass at the origin (i.e., $m = 1$), by the solid right circular cone of uniform mass density σ (g/cm^3) shown in Fig. 5. Since the cone is not a point mass, we need to find the forces induced by the individual mass elements that constitute the cone and to sum them by integration. With M changed to $\sigma\, dV$ and $x_0 = y_0 = z_0 = 0$, (20) and (21) give

$$\mathbf{F}(0,0,0) = G \iiint\limits_{\mathcal{V}} \frac{x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}}{(x^2+y^2+z^2)^{3/2}}\, \sigma\, dV. \tag{22}$$

Surely, the symmetry about the z axis implies that the x and y components of $\mathbf{F}$ will turn out to be zero so that (22) can be simplified to

$$\mathbf{F}(0,0,0) = G\sigma\hat{\mathbf{k}} \iiint\limits_{\mathcal{V}} \frac{z}{(x^2+y^2+z^2)^{3/2}}\, dV. \tag{23}$$

In what coordinate system shall we work out the integral? That is, how shall we parametrize the conical region? Spherical coordinates seem like a good choice inasmuch as the conical surface is a constant-ϕ coordinate surface. However, the top surface, $z = h$, is *not* a constant-coordinate surface. That is, $\rho \neq$ constant, $\phi \neq$ constant, and $\theta \neq$ constant on $z = h$. Using cylindrical coordinates the top surface *is* a constant-coordinate surface (namely, $z = h$) but the conical surface is not. Thus, the choice seems to be "six of one, half a dozen of the other." Let us work the problem both ways.

Using cylindrical coordinates: That is, choose u, v, w to be the familiar cylindrical coordinates r, θ, z, respectively. Then $dV = r\, dr\, d\theta\, dz$ and $x^2 + y^2 = r^2$ so (23) becomes

$$\begin{aligned}
\mathbf{F} &= \sigma G\hat{\mathbf{k}} \int_0^h \int_0^{2\pi} \int_0^{z\tan\alpha} \frac{zr\, dr\, d\theta\, dz}{(r^2+z^2)^{3/2}} \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h \int_{u=z^2}^{u=z^2(1+\tan^2\alpha)} \frac{z\, du\, dz}{2u^{3/2}} \qquad (u = r^2 + z^2) \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h z\left(-\frac{1}{\sqrt{u}}\right)\Bigg|_{z^2}^{z^2/\cos^2\alpha} dz \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h z\left(-\frac{\cos\alpha}{z} + \frac{1}{z}\right) dz,
\end{aligned} \tag{24}$$

so that

$$\mathbf{F} = 2\pi\sigma G h(1 - \cos\alpha)\hat{\mathbf{k}}. \tag{25}$$

In case the $z\tan\alpha$ limit was not clear, note that $r/z = \tan\alpha$ on the conical surface.

Using spherical coordinates: This time choose u, v, w to be the spherical coordinates ρ, ϕ, θ, respectively. Then $dV = \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta$, $z = \rho\cos\phi$, and $x^2 + y^2 + z^2 = \rho^2$ so (23) becomes

$$\begin{aligned}
\mathbf{F} &= \sigma G\hat{\mathbf{k}} \int_0^{2\pi} \int_0^{\alpha} \int_0^{h/\cos\phi} \frac{\rho\cos\phi}{(\rho^2)^{3/2}}\, \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^{2\pi} \cos\phi \sin\phi \frac{h}{\cos\phi}\, d\phi = 2\pi\sigma G h(1 - \cos\alpha)\hat{\mathbf{k}},
\end{aligned} \tag{26}$$

which result agrees with (25).

COMMENT. As a partial check, observe from (26) that $\mathbf{F} \to \mathbf{0}$ as $\alpha \to 0$, as it should. ▮

Closure. With a three-dimensional region generated by a three-parameter family $x = x(u, v, w)$, $y = y(u, v, w)$, $z = z(u, v, w)$, and the position vector given by $\mathbf{R} = x = x(u, v, w)\hat{\mathbf{i}} + y(u, v, w)\hat{\mathbf{j}} + z(u, v, w)\hat{\mathbf{k}}$, the volume element dV is defined as the volume of the parallelipiped with edges $\mathbf{R}_u\, du$, $\mathbf{R}_v\, dv$, $\mathbf{R}_w\, dw$ as shown in Fig. 1. Thus,

$$\begin{aligned} dV &= |(\mathbf{R}_u\, du) \cdot (\mathbf{R}_v\, dv) \times (\mathbf{R}_w\, dw)| \\ &= |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\, du\, dv\, dw \\ &= \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\, dv\, dw \\ &= \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\, dv\, dw. \end{aligned}$$

That is, dV is the absolute magnitude of the Jacobian of x, y, z with respect to u, v, w times $du\, dv\, dw$.

EXERCISES 15.6

1. Compute the volume of a sphere of radius R by triple integration:

(a) using cylindrical coordinates
(b) using spherical coordinates.

2. Show by triple integration that the volume of the right circular cone shown in Fig. 5 is $V = (\pi h^3/3) \tan^2 \alpha$:

(a) using cylindrical coordinates
(b) using spherical coordinates.

3. Evaluate the gravitational force $\mathbf{F}$ given by (23), this time letting $\mathcal{V}$ be the hemisphere $x^2 + y^2 + z^2 \le R^2$, for $0 \le z \le R$

(a) using cylindrical coordinates
(b) using spherical coordinates.

4. (*Mass and center of gravity*) First, read Exercise 4 of Section 15.3. There, the mass was a two-dimensional distribution over a region $\mathcal{R}$ in the x, y plane. Here, we consider a distribution over a three-dimensional region $\mathcal{R}$ in x, y, z space. If the mass density is $\sigma(x, y, z)$ [i.e., $\sigma(x, y, z)$ is the mass per unit volume at x, y, z], then the x, y, z coordinates of the **center of gravity** are defined as

$$\begin{aligned} x_c &= \frac{1}{M} \iiint_{\mathcal{R}} x\sigma(x, y, z)\, dV, \\ y_c &= \frac{1}{M} \iiint_{\mathcal{R}} y\sigma(x, y, z)\, dV, \\ z_c &= \frac{1}{M} \iiint_{\mathcal{R}} z\sigma(x, y, z)\, dV, \end{aligned} \tag{4.1}$$

where

$$M = \iiint_{\mathcal{R}} \sigma(x, y, z)\, dV$$

is the total mass. In each of the following, take the density σ to be a constant.

(a) Let $\mathcal{R}$ be the solid cone considered in Example 3. Compute z_c.
(b) Let $\mathcal{R}$ be one eighth of the sphere $x^2 + y^2 + z^2 \le R^2$, namely, that part which lies within $x \ge 0$, $y \ge 0$, $z \ge 0$. Compute x_c, y_c, z_c.
(c) Let $\mathcal{R}$ be that part of the sphere $x^2 + y^2 + z^2 \le R^2$

which lies between $z = a$ and $z = R$ (where $0 < a < R$). Compute z_c.
(d) Let $\mathcal{R}$ be the spherical sector $0 \le \rho \le R$, $0 \le \phi \le \alpha$, $0 \le \theta < 2\pi$. Compute z_c.

5. (*Moments of inertia*) First, read Exercise 13 of Section 15.3. In each of the following, take the density σ to be a constant. Let M be the total mass.

(a) Consider a hollow sphere $a^2 \le x^2 + y^2 + z^2 \le b^2$. Show that

$$I_x = \frac{2}{5} M \frac{b^5 - a^5}{b^3 - a^3}.$$

(b) In the limit $a \to b$, the hollow sphere in part (a) is called a "thin shell." For a thin shell of radius b, show that the answer to part (a) reduces to the form $I_x = \frac{2}{3}Mb^2$.
(c) Consider a hollow right circular cylinder $a^2 \le x^2 + y^2 \le b^2, 0 \le z \le h$. Show that $I_x = (M/12)(3b^2 + 3a^2 + 4h^2)$.
(d) For the cylinder in part (c), show that $I_z = (M/2)(a^2+b^2)$.
(e) For the right circular cone shown in Fig. 5, show that $I_z = (3M/10)(h \tan \alpha)^2$.
(f) For the right circular cone shown in Fig. 5, show that $I_x = (3Mh^2/20)(\tan^2 \alpha + 4)$.

6. In the case of cylindrical coordinates, we set $u = r$, $v = \theta$, $w = z$. Show that the result (8) would obtain if, instead, we set $u = \theta$, $v = r$, $w = z$, say.

7. (*Gravitational force induced by hollow sphere*) Consider a hollow sphere, $a^2 \le x^2 + y^2 + z^2 \le b^2$, of uniform mass density σ.

(a) Show that the gravitational force of attraction induced by the hollow sphere, per unit mass at $(0, 0, z_0)$, is

$$\mathbf{F}(0,0,z_0) = \sigma G \hat{\mathbf{k}} \iiint_{\mathcal{V}} \frac{(z - z_0)dV}{[x^2 + y^2 + (z - z_0)^2]^{3/2}}. \tag{7.1}$$

(b) Carrying out the integration in (7.1), show that

$$\mathbf{F}(0,0,z_0) = \begin{cases} -\dfrac{MG}{z_0^2}\hat{\mathbf{k}} & \text{if } |z_0| > b \\ 0 & \text{if } |z_0| < a. \end{cases} \tag{7.2}$$

NOTE: The result (7.2) is remarkable for it says that the force induced by the hollow sphere at any point outside the sphere is the same as if the entire mass of the sphere were compressed into a *point* mass M at the origin, and that there is *no* force within the cavity.
(c) Evaluate $\mathbf{F}(0, 0, z_0)$ for the case where $a < z_0 < b$. HINT: The result obtained in part (b) should be all that you need.

8. Using the results obtained in Example 3:

(a) What is the gravitational force induced by an infinite slab ($-\infty < x < \infty$, $-\infty < y < \infty$, $0 < z < h$) of uniform mass density σ, per unit mass, at any point on either face of the slab ($z = 0$ or $z = h$)?
(b) What is the gravitational force induced by an infinite half-space ($-\infty < x < \infty$, $-\infty < y < \infty$, $0 \le z < \infty$) of uniform mass density σ, per unit mass, at any point on the face $z = 0$?

Chapter 15 Review

The single most important idea about curves, surfaces, and volumes is their parametrization by

$$\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}, \tag{1a}$$
$$\mathbf{R}(u, v) = x(u, v)\hat{\mathbf{i}} + y(u, v)\hat{\mathbf{j}} + z(u, v)\hat{\mathbf{k}}, \tag{1b}$$
$$\mathbf{R}(u, v, w) = x(u, v, w)\hat{\mathbf{i}} + y(u, v, w)\hat{\mathbf{j}} + z(u, v, w)\hat{\mathbf{k}}, \tag{1c}$$

respectively, where $\mathbf{R}$ is the position vector. We could have used u instead of τ in (1a), to promote the (u), (u, v), (u, v, w) pattern in (1a)–(1c), but τ is more traditional for curves. As should not be surprising, all quantities of interest, for curves, surfaces, and volumes can be expressed in terms of $\mathbf{R}$. The differential arc length, area element, and volume element are

$$ds = \sqrt{\mathbf{R}'(\tau) \cdot \mathbf{R}'(\tau)}\, d\tau, \tag{2}$$

$$dA = \|\mathbf{R}_u \times \mathbf{R}_v\|\, du\, dv, \tag{3}$$

$$dV = |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\, du\, dv\, dw, \tag{4}$$

respectively, and the normal to a surface is, to within a factor of ± 1,

$$\hat{\mathrm{n}} = \frac{\mathbf{R}_u \times \mathbf{R}_v}{\|\mathbf{R}_u \times \mathbf{R}_v\|}. \tag{5}$$

Computationally, it is important to express these results in terms of the components x, y, z of $\mathbf{R}$. Accordingly, (2)–(4) become

$$ds = \sqrt{x'^2 + y'^2 + z'^2}\, d\tau, \tag{6}$$

$$\begin{aligned} dA &= \sqrt{EG - F^2}\, du\, dv, \\ &\quad E = x_u^2 + y_u^2 + z_u^2 \\ &\quad F = x_u x_v + y_u y_v + z_u z_v \\ &\quad G = x_v^2 + y_v^2 + z_v^2, \end{aligned} \tag{7}$$

$$dV = \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\, dv\, dw = \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\, dv\, dw. \tag{8}$$

We note two special cases of (7), one when the surface is flat and lies in the x, y plane ($z = 0$), and one when the surface is known in the form $z = f(x, y)$.

$$z = 0: \qquad dA = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv, \tag{9a}$$

$$z = f(x, y): \qquad dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy. \tag{9b}$$

Observe the appearance in (8) and (9a) of the Jacobian determinants.

Chapter 16

Scalar and Vector Field Theory

16.1 Introduction

A great many phenomena are governed by ordinary differential equations (ODE's). That is, there is only one independent variable, such as a space variable or time, and one or more dependent variables. If there are two or more independent variables, then the dependent variable is generally called a *field*, and the governing differential equation will be a partial differential equation (PDE), known as a *field equation*. For instance, the concentration $c(x, y, z, t)$ of a pollutant that spreads by diffusion in a lake is governed by the PDE

$$\alpha^2 \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2} \right) = \frac{\partial c}{\partial t}, \tag{1}$$

where α^2 is a constant known as the diffusivity. As another example, let $\sigma(x, y, z, t)$ and $\mathbf{v}(x, y, z, t)$ be the mass density and velocity of a certain fluid flow, respectively. Then the principle of conservation of mass implies that σ and $\mathbf{v}$ are related through the PDE

$$\frac{\partial \sigma}{\partial t} + \frac{\partial}{\partial x}(\sigma v_x) + \frac{\partial}{\partial y}(\sigma v_y) + \frac{\partial}{\partial z}(\sigma v_z) = 0, \tag{2}$$

where v_x, v_y, v_z are the x, y, z components of $\mathbf{v}$, respectively. In this chapter we study the calculus of such scalar and vector fields.

Our plan is as follows. In Sections 16.3–16.6 we introduce the so-called divergence, gradient, and curl, which are specific differential operators that can act on scalar and vector fields. Those results are extended to non-Cartesian coordinates (cylindrical and spherical) in Section 16.7, which is optional so it can be omitted in a shorter course.

The "payoff" comes mostly in Sections 16.8–16.10, on the Gauss divergence theorem, Stokes's theorem, and irrotational fields. For instance, in Section 16.8 we derive (1) and (2) above, using the divergence theorem. Thus, one of the major outcomes of this chapter is the derivation of some of the important field equations of mathematical physics. After pausing in Chapter 17 to develop the Fourier series and Fourier transform, we will then come back to those field equations and learn how to solve them – in Chapters 18–20 on PDE's. We continue to treat the slightly more difficult non-Cartesian cases as optional, in Sections 16.8–16.10, by including them only in optional subsections.

16.2 Preliminaries

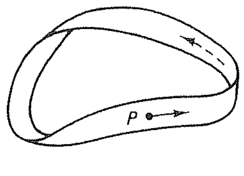


Figure 1. Möbius band.

16.2.1. Topological considerations. Throughout this chapter we ask that the curves, surfaces, and volumes under consideration be sufficiently "decent" for our purposes. Specifically, we ask the curves to be piecewise smooth (Section 15.2.1). We ask the surfaces to be piecewise smooth too (Section 15.3.2) but, in addition, we ask them to be orientable. Since nonorientable surfaces are rarely encountered in applications, and since a rigorous definition would be a long story, we will define the term only in an intuitive way. Namely, we say that a surface is **orientable** if it is two-sided, that is, if we can paint one side green and the other side blue. The classic example of a *non*orientable surface is the *Möbius band*, which can be constructed by taking a rectangular strip of paper, twisting one end through 180°, then taping that end to the other (Fig. 1). If an insect starts at any point P on one side of the band and follow the arrows, it eventually visits the entire surface, so that the Möbius band has only one side. Our objection to such a surface is that we are going to "orient" a given surface $\mathcal{S}$ by defining a normal vector $\hat{\mathbf{n}}$ at some point on $\mathcal{S}$ and then extending the field of normals continuously over the rest of $\mathcal{S}$. But when we do that with the Möbius band, say, we end up with *two* vectors, $\hat{\mathbf{n}}$ and $-\hat{\mathbf{n}}$, at each point on $\mathcal{S}$.

Finally, we ask that the volumes under consideration be regions (Section 13.2.2) that are bounded by piecewise smooth orientable surfaces.

16.2.2. Scalar and vector fields. The objects of interest in this chapter are scalar fields and vector fields. By a **scalar field** we shall mean a *scalar valued function*, say f, defined over a region $\mathcal{R}$, which is a one-, two-, or three-dimensional connected subset of 3-space. Besides the space variables, f may depend on the time t as well. Normally, f will be real valued.

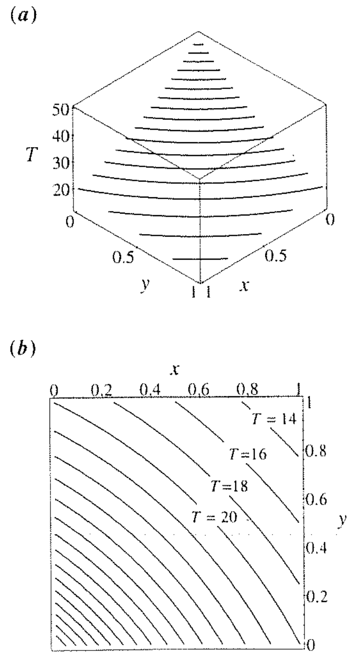


Figure 2. Plots of $T(x, y)$.

EXAMPLE 1. *Temperature Field.* Let $T(x, y)$ be the temperature at any given point (x, y) in a square plate which extends over $0 \leq x \leq 1$, $0 \leq y \leq 1$, and let

$$T(x, y) = \frac{100}{(x+1)^2 + (y+1)^2}. \tag{1}$$

In this case the region $\mathcal{R}$ is the two-dimensional set $0 \leq x \leq 1$, $0 \leq y \leq 1$, and the scalar function is $T(x, y)$, given by (1). By way of graphical display, we have used *Maple* to obtain the three-dimensional plot of T, above the x, y plane, that is given in Fig. 2a. Plotted are a number of **level curves** of T, curves along which $T(x, y)$ is a constant, like level roads on a mountain. The projection of those curves down onto the x, y plane give the level curves shown in Fig. 2b. We can see from (1) that the level curves of T, in the x, y plane, are the concentric circles $(x+1)^2 + (y+1)^2 = \text{constant}$, centered at $(-1, -1)$. ▮

In Example 1 we encountered level curves. More generally, we say that the **level set** of value c, of a function $f(x_1, \ldots, x_n)$, is the set of points $(x_1, \ldots, x_n)$ for which $f(x_1, \ldots, x_n) = c$. If $n = 2$, the level set is, in general, a level *curve*, and if $n = 3$ the level set is, in general, a level *surface*. To see why we say "in general," consider a few examples for the case where $n = 2$. Let

$$f(x, y) = \ln(x^2 + y^2), \qquad g(x, y) = \sin(x + 3y), \qquad h(x, y) = 6,$$

each defined over the whole plane. It is not hard to see that the level sets of f are concentric circles centered at the origin. Specifically, for any given value c ($-\infty < c < \infty$) the level set of f is the circle $x^2 + y^2 = e^c$. Turning to g, the level set of value $\frac{1}{2}$, say, is not just a single curve, it is the *set* of straight lines $x + 3y = \pi/6 + 2n\pi$ ($n = 0, \pm 1, \pm 2, \ldots$), and the level set of value 4, say, is *empty*. Finally, turning to h, the level set of value 2, say, is empty, whereas the level set of value 6 is the whole plane!

Just as a scalar field is a scalar valued function defined over a given region, a *vector valued function* defined over a given region is called a **vector field**.*

EXAMPLE 2. *Fluid Velocity Field.* Imagine a steady, uniform flow of fluid (such as air or water), at a speed U meters/second, parallel to a wall as shown in Fig. 3. That is, at every instant and at every point (x, y) in the region (namely, $-\infty < x < \infty, 0 < y < \infty$), the fluid velocity $\mathbf{v}$ is equal to $U\hat{\mathbf{i}}$, where U is a constant. Then the vector function

$$\mathbf{v} = U\hat{\mathbf{i}}, \tag{2}$$

in the stated region, constitutes a vector field – a simple one, but a vector field nonetheless.

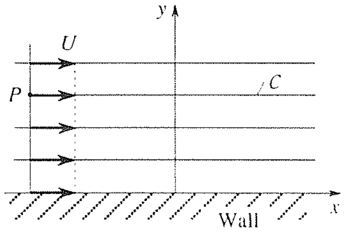


Figure 3. Steady, uniform flow field.

Since the flow is steady, every fluid particle that passes through the point P (Fig. 3), at one time or another, must follow the same path, namely, the curve C. Fluid mechanicists call such curves *streamlines*. (In general, streamlines do not exist if the flow is *un*steady.) Thus we speak of the family of horizontal lines in Fig. 3 as the *streamline pattern*.

As a more complicated case, let the wall have a semicircular bump of radius a, centered at the origin, and let the fluid velocity field be

$$\mathbf{v} = U\hat{\mathbf{i}} + \frac{Ua^2}{(x^2+y^2)^2}\left[(y^2 - x^2)\hat{\mathbf{i}} - 2xy\hat{\mathbf{j}}\right], \qquad (a^2 \le x^2 + y^2 < \infty) \tag{3}$$

as depicted in Fig. 4. At the top of the bump, for example, where $x = 0$ and $y = a$,

$$\mathbf{v}(0, a) = U\hat{\mathbf{i}} + \frac{Ua^2}{a^4}(a^2\hat{\mathbf{i}} - 0\hat{\mathbf{j}}) = 2U\hat{\mathbf{i}}.$$

Furthermore, it follows from (3) that $\mathbf{v} = 0$ at $(\pm a, 0)$, which points are therefore called *stagnation points* of the flow.

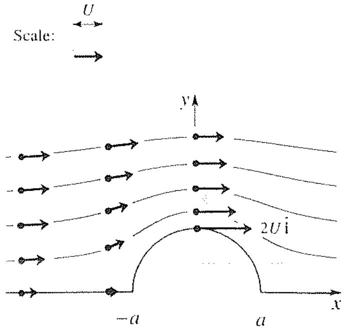


Figure 4. The flow field (3).

*That is, a vector field $\mathbf{v}$ on a region $\mathcal{R}$, which is a subset of $\mathbb{R}^n$, is a transformation or mapping $\mathbf{v} : \mathcal{R} \subset \mathbb{R}^n \to \mathbb{R}^n$. In this chapter n will be 2 or 3, the general case is considered in J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Francisco: W. H. Freeman, 1976). Similarly, a scalar field f on a region $\mathcal{R}$, which is a subset of $\mathbb{R}^n$, is a mapping $f : \mathcal{R} \subset \mathbb{R}^n \to \mathbb{R}^1$.

COMMENT. In Section 16.10 we derive the PDE governing the flow field shown in Fig. 4. Later, when we study PDE's, we will solve that PDE and obtain the result stated in (3). In fact, we will solve the same problem again, in Section 23.6, using the method of conformal mapping. ▮

Closure. Scalar and vector fields are scalar and vector valued *functions*, respectively, defined on one-, two-, or three-dimensional regions of space. They may be constant or nonconstant, and they may or may not vary with the time t as well as the space variables. Examples of scalar fields are the temperature fields $T(x, y, z) = 100(x^2 + y^2 + 5z^2)$ and $T(x, t) = 50(\sin x)e^{-2t}$, and examples of vector fields are the fluid velocity fields $\mathbf{v}(x, y) = 4xy\hat{\mathbf{i}} - (x + 3y^2)\hat{\mathbf{j}}$ and $\mathbf{v}(x, y, z, t) = 10\hat{\mathbf{i}} + x^3yz\sin t\hat{\mathbf{k}}$.

Computer software. The computer plots in Fig. 2 were obtained using the *Maple* **plot3d** and **contourplot** commands, as follows. For Fig. 2a we used

```
with(plots):
plot3d(100/((x + 1)^2 + (y + 1)^2), x = 0..1, y = 0..1, style = contour);
```

and for Fig. 2b we used

```
with(plots):
contourplot(100/((x + 1)^2 + (y + 1)^2), x = 0..1, y = 0..1);
```

If we omit the style = contour option in the first, then instead of the plot being comprised of level curves it would, as the default condition, be comprised of constant-x and constant-y curves. That is, the display would be the u, v coordinate curves under the parametrization $x = u$ and $y = v$.

EXERCISES 16.2

1. Let $T(x, y)$ be a temperature field in $|x| \leq 10$, $0 \leq y \leq 10$. Determine, and sketch the $T = 0, 20, 40, 60$ level sets. If the level set is empty, state that.

(a) $T = x^2 + y^2$	(b) $T = 10(x - y)$
(c) $T = 10(x + y)$	(d) $T = x^2 - y^2$
(e) $T = 2x^2$	(f) $T = 100/(x^2 + y^2 + 1)$
(g) $T = x + 20y$	(h) $T = x^2y^2/20$
(i) $T = 60\sin(\pi xy/128)$	(j) $T = 53$
(k) $T = \sin[(x^2 + y^2)/24]$	(l) $T = 20$

2. The position vector $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$ to any point (x, y) in the x, y plane constitutes a vector field. Draw the $\mathbf{R}$ vector (to scale) at eight points that are equally spaced around the circle $r = 1$, and at eight points that are equally spaced around the circle $r = 3$.

3. Consider the vector field $\mathbf{v} = x\hat{\mathbf{i}} - y\hat{\mathbf{j}}$ in $0 \leq x < \infty$, $0 \leq y < \infty$.

(a) On a single graph, sketch the $\mathbf{v}$ vectors at the 16 points (m, n), where m and n are integers such that $0 \leq m \leq 3$ and $0 \leq n \leq 3$. (Here it will help to assign a convenient scale, as we have done in Fig. 4, so that the arrow representations of the vectors will not lie on top of each other.)

(b) Determine the curves along which $\mathbf{v}$ has constant magnitude and those along which $\mathbf{v}$ has constant direction. Are these results in agreement with your sketch of the field in part (a)?

4. Given the vector field $\mathbf{w}(x, y, z)$, determine parametric equations $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$ for the curve(s) through the point $(2, 3, -1)$ along which $\mathbf{w}$ has constant direction.

(a) $\mathbf{w} = y\hat{\mathbf{i}} - x\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(b) $\mathbf{w} = (y + 2z)\hat{\mathbf{i}} - x\hat{\mathbf{j}} + (x + y)\hat{\mathbf{k}}$
(c) $\mathbf{w} = x^2\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(d) $\mathbf{w} = yz\hat{\mathbf{i}} - \hat{\mathbf{j}} - (x + 2)\hat{\mathbf{k}}$
(e) $\mathbf{w} = z^2\hat{\mathbf{i}} - 3\hat{\mathbf{j}} - (y - x)\hat{\mathbf{k}}$
(f) $\mathbf{w} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$

5. (a) For the velocity field defined by (3), determine the *speed* $\|\mathbf{v}\|$ as a function of x, along the wall (namely, $y = 0$ for $|x| \geq a$ and $x^2 + y^2 = a^2$ for $|x| \leq a$).
(b) Show that as $r = \sqrt{x^2 + y^2} \to \infty$ along any ray $y = mx$, the velocity $\mathbf{v}$ tends to the uniform flow $U\hat{\mathbf{i}}$, which flow is called the *free stream*.

6. Recall the plane polar coordinate base vectors $\hat{\mathbf{e}}_r(\theta)$ and $\hat{\mathbf{e}}_\theta(\theta)$ from Section 14.6. Each of these is a vector field because there is an $\hat{\mathbf{e}}_r$ vector and an $\hat{\mathbf{e}}_\theta$ vector defined at each point in the plane (except the origin).

(a) Sketch the $\hat{\mathbf{e}}_r(\theta)$ field.
(b) Sketch the $\hat{\mathbf{e}}_\theta(\theta)$ field.

7. Use computer software to obtain plots of the four streamlines shown in Fig. 4, taking $U = a = 1$ and taking their initial points as $(-3.5, 0.5)$, $(-3.5, 1)$, $(-3.5, 1.5)$, and $(-3.5, 2)$. HINT: Equation (3) gives us the ODEs

$$\begin{aligned} x'(t) &= 1 + \frac{y^2 - x^2}{(x^2 + y^2)^2}, \\ y'(t) &= -\frac{2xy}{(x^2 + y^2)^2} \end{aligned} \tag{7.1}$$

for the motion $x(t), y(t)$ of any given fluid particle.

16.3 Divergence

Fundamental in scalar and vector field theory are certain differential operators, known as the divergence, gradient, and curl, that operate on those fields. We will introduce the divergence in this section and the gradient and curl in the next two.

Let $\mathbf{v}$ be a vector field defined in a region $\mathcal{R}$. The physical nature of $\mathbf{v}$ is immaterial here, but for definiteness let us think of $\mathbf{v}$ as a fluid velocity field. Focusing our attention on a particular point $P = (x, y, z)$ in the flow, let us introduce a **control volume** around P, as shown in Fig. 1, and denote it as $\mathcal{B}$. For example, $\mathcal{B}$ might be chosen to be a prism, a sphere, or an arbitrary "potato" shape as in the figure. A control volume is only a mathematical region rather than a physical presence, and it is normally introduced so that we can keep track of the flux of some quantity of interest such as mass, electric charge, or heat.

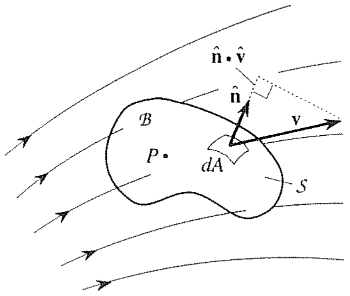


Figure 1. Control volume.

For instance, consider the integral*

$$\boxed{I = \int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA,} \tag{1}$$

*In Chapter 15 we denote double and triple integrals by $\iint$ and $\iiint$, respectively. In this section we switch to the more compact $\int$ notation used in (1) unless, of course, integration limits are to be spelled out. Another point of notation is that some authors prefer to combine $\hat{\mathbf{n}}$ and dA in (1) as $\hat{\mathbf{n}}\, dA = d\mathbf{A}$; then (1) becomes $\int_S \mathbf{v} \cdot d\mathbf{A}$, which is a bit more compact than (1), but we will not use that notation.

where $\hat{\mathbf{n}}$ is a unit outward normal vector at each point on $\mathcal{S}$, as illustrated, at a representative point in Fig. 1. The set of all such $\hat{\mathbf{n}}$ vectors on $\mathcal{S}$ would look something like the quills on a frightened porcupine. What is the physical significance of I? If, at a given point on the surface $\mathcal{S}$ of $\mathcal{B}$, we break the fluid velocity $\mathbf{v}$ into normal and tangential components (Fig. 2), then the flux across dA is due entirely to the normal component $\hat{\mathbf{n}} \cdot \mathbf{v}$ since the tangential flow is along the surface and therefore does not cross it. Then the outflow through the surface element dA is the normal velocity component $\hat{\mathbf{n}} \cdot \mathbf{v}$ (meters per second, say) times dA (square meters), that is, $\hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ m^3/sec. Thus, integrating over the entire surface $\mathcal{S}$, I is the *net outflow*, in m^3/sec, say, across $\mathcal{S}$ – outflow because we took $\hat{\mathbf{n}}$ to be the outward unit normal. This outflow can be positive, zero, or negative. For example, if heat is extracted from the fluid and the fluid contracts, one can imagine a net *in*flow into the control volume so that the outflow will be negative.

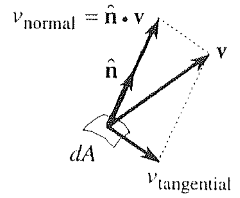


Figure 2. Flux across dA.

Thus, I is an example of a flux integral. We will call it a **volume flux** because it is volume per unit time. If, besides denoting the vector velocity field as $\mathbf{v}(x, y, z, t)$, we also denote the scalar density field as $\sigma(x, y, z, t)$ (mass per unit volume), then

$$J = \int_{\mathcal{S}} \sigma \hat{\mathbf{n}} \cdot \mathbf{v}\, dA \tag{2}$$

would also be a flux, not a volume flux but a mass flux because the dimensions of $\sigma \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ are mass per unit volume times volume per unit time, hence mass per unit time.

Next, let us divide the volume flux integral in (1) by the volume V of $\mathcal{B}$ to obtain the outflow *per unit volume*. Finally, we shrink $\mathcal{B}$ down to point P and obtain the outflow per unit volume at the point P. This result is called the **divergence** of $\mathbf{v}$ at P and is defined as

$$\boxed{\operatorname{div} \mathbf{v}(P) \equiv \lim_{\mathcal{B} \to 0} \left\{ \frac{\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\},} \tag{3}$$

where $\mathcal{B} \to 0$ means that $\mathcal{B}$ shrinks to the point P in such a way that the maximum linear dimension (the "diameter") of $\mathcal{B}$ tends to zero.* If we assume that $\mathbf{v}$ is C^1 and that $\mathcal{B}$ has a piecewise smooth orientable surface $\mathcal{S}$, then it can be shown that the limit in (3) does indeed exist at each point P in the field.†

Observe that $\operatorname{div} \mathbf{v}(P)$ is a scalar at each point P since $\hat{\mathbf{n}} \cdot \mathbf{v}$, dA, and V are scalars. Thus, $\operatorname{div} \mathbf{v}$ is itself a scalar field associated with the given vector field $\mathbf{v}$.

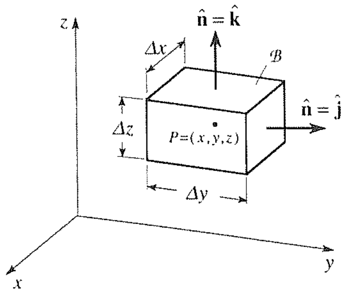
Observe, further, that (3) provides an intrinsic, or invariant, definition of $\operatorname{div} \mathbf{v}$. That is, it contains no reference to any particular coordinate system so the value of

*Observe that the limit as $\mathcal{B} \to 0$ is not the same as the limit as $V \to 0$. For instance, if we make a coin thinner and thinner, then its volume V tends to zero even though it is not true that $\mathcal{B} \to 0$.

†Let $\mathbf{v} = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$, where t is the time. By $\mathbf{v}$ being C^1 in $\mathcal{R}$, we mean that v_x, v_y and v_z are all C^1 in $\mathcal{R}$; i.e., $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ of v_x, v_y, and v_z all exist and are continuous in $\mathcal{R}$. Since all the first-order partial derivatives of $\mathbf{v}$ exist and are continuous in $\mathcal{R}$, we also say that $\mathbf{v}$ is **continuously differentiable** in $\mathcal{R}$.

div $\mathbf{v}(P)$ at any given point P is uniquely determined, independent of the choice of the reference coordinate system.

Thus, the definition of div $\mathbf{v}$ given by (3) has two advantages: it readily admits a clear and simple physical interpretation of div $\mathbf{v}(P)$ as the outflow per unit volume at P, and it is invariant with respect to coordinate system. However, the limit definition (3) is of little use computationally. To obtain a computationally convenient expression, we need to introduce a reference coordinate system. Introducing a Cartesian system, since that is evidently the simplest choice, let us carry out the limit indicated in (3). Since the limit in (3) is independent of the shape of $\mathcal{B}$, let us choose the simplest shape, namely, one bounded by constant coordinate surfaces. Thus, for $\mathcal{B}$ let us choose the rectangular prism shown in Fig. 3, with $P = (x, y, z)$ at its center.



z
$\hat{\mathbf{n}} = \hat{\mathbf{k}}$
Δx
$\mathcal{B}$
$\hat{\mathbf{n}} = \hat{\mathbf{j}}$
Δz
$P=(x,y,z)$
Δy
x
y

Figure 3. Cartesian coordinates.

Consider first the contribution from the front and back faces, on which the outward unit normal is $\hat{\mathbf{n}} = +\hat{\mathbf{i}}$ and $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$, respectively. Let us express $\mathbf{v} = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$, where the x, y, z subscripts specify the x, y, z components of $\mathbf{v}$ rather than partial derivatives,* and where we have not included any time dependence in $\mathbf{v}$ because it will not play an active role in the present discussion. Then

$$\begin{aligned}\int_{\text{front face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{\text{front face}} \hat{\mathbf{i}} \cdot \left(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}\right) dA \\ &= \int_{\text{front face}} v_x\left(x + \frac{\Delta x}{2}, y', z'\right) dy'\, dz' \\ &= v_x\left(x + \frac{\Delta x}{2}, y_1, z_1\right) \Delta y\, \Delta z, \end{aligned} \tag{4}$$

for some point $(x + \Delta x/2, y_1, z_1)$ on the front face, where the last equality in (4) follows from the mean value theorem (4c) in Section 15.3. [To distinguish the x, y, z coordinates of the fixed point $P = (x, y, z)$ from the dummy variables of integration, we use x', y', z' for the latter.] Similarly,

$$\begin{aligned}\int_{\text{back face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{\text{back face}} -\hat{\mathbf{i}} \cdot \left(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}\right) dA \\ &= -v_x\left(x - \frac{\Delta x}{2}, y_2, z_2\right) \Delta y\, \Delta z \end{aligned} \tag{5}$$

for some point $(x - \Delta x/2, y_2, z_2)$ on the back face. Since the volume of $\mathcal{B}$ is $V = \Delta x\, \Delta y\, \Delta z$, it follows that

$$\begin{aligned}&\lim_{\mathcal{B}\to 0}\left\{\frac{\int_{\text{front face+back face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V}\right\} \\ &\quad= \lim_{\Delta x, \Delta y, \Delta z \to 0}\left\{\frac{\left[v_x\left(x + \Delta x/2, y, z\right) - v_x\left(x - \Delta x/2, y, z\right)\right] \Delta y \Delta z}{\Delta x\, \Delta y\, \Delta z}\right\} \\ &\quad= \lim_{\Delta x \to 0} \frac{v_x\left(x + \Delta x/2, y, z\right) - v_x\left(x - \Delta x/2, y, z\right)}{\Delta x} = \frac{\partial v_x}{\partial x}. \end{aligned} \tag{6}$$

*Thus, partial derivatives will be expressed as $\partial(\ )/\partial x$, $\partial(\ )/\partial y$, and so on.

Similarly, the top and bottom faces contribute $\partial v_z/\partial z$, and the right and left faces contribute $\partial v_y/\partial y$, so that

$$\boxed{\operatorname{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}.} \tag{7}$$

Analogous expressions are given in Section 16.7, for cylindrical and spherical coordinates.

EXAMPLE 1. Evaluate the divergence of the field $\mathbf{v} = (x^2/z)\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + yz\hat{\mathbf{k}}$ at the point $P = (2, 9, -1)$. First, let us check to see if $\mathbf{v}$ is C^1 in some region $\mathcal{R}$ containing P. The partial derivatives of $v_x = x^2/z$, $v_y = -3$ and $v_z = yz$ with respect to x, y, and z exist and are continuous functions of x, y, z everywhere except on the plane $z = 0$, where $\partial v_x/\partial x = 2x/z$ and $\partial v_x/\partial z = -x^2/z^2$ are undefined ("blow up"). But P is not in the plane $z = 0$ so we can say that $\mathbf{v}$ is indeed C^1 in a region $\mathcal{R}$ containing P, where $\mathcal{R}$ is the half-space $-\infty < x < \infty$, $-\infty < y < \infty$, $-\infty < z < 0$. Then (7) gives

$$\operatorname{div}\mathbf{v} = \frac{\partial}{\partial x}\left(\frac{x^2}{z}\right) + \frac{\partial}{\partial y}(-3) + \frac{\partial}{\partial z}(yz) = \frac{2x}{z} + y,$$

so at P we have $\operatorname{div}\mathbf{v} = -4 + 9 = 5$.

COMMENT. For brevity, in subsequent examples we will not verify explicitly that $\mathbf{v}$ is C^1. ▮

EXAMPLE 2. If $\mathbf{v} = y^3zt\hat{\mathbf{j}}$, where t is the time, then

$$\operatorname{div}\mathbf{v} = \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}(y^3zt) + \frac{\partial}{\partial z}(0) = 3y^2zt.$$

That is, t is treated as a constant because only space derivatives are involved. ▮

Rewriting (7) as

$$\operatorname{div}\mathbf{v} = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \tag{8}$$

suggests that we define a vector differential operator

$$\boxed{\nabla \equiv \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z},} \tag{9}$$

in which case we can write div v in the operator form

$$\boxed{\operatorname{div}\mathbf{v} = \nabla \cdot \mathbf{v}.} \tag{10}$$

We will have more to say about the $\nabla \cdot \mathbf{v}$ operation in subsequent sections, but one point that must be mentioned immediately is that $\nabla \cdot \mathbf{v}$ is not a dot product in the usual sense. Specifically, it is *not* true that $\nabla \cdot \mathbf{v}$ equals $\|\nabla\|$ times $\|\mathbf{v}\|$ times the cosine of the angle between ∇ and $\mathbf{v}$ because $\nabla = \hat{\mathbf{i}}(\partial/\partial x) + \hat{\mathbf{j}}(\partial/\partial y) + \hat{\mathbf{k}}(\partial/\partial z)$ is a vector *differential operator*, not a vector with a length and a direction. Thus, $\|\nabla\|$ is not defined nor is the "angle between ∇ and $\mathbf{v}$." Furtheremore, $\nabla \cdot \mathbf{v}$ is *not* the same as $\mathbf{v} \cdot \nabla$ since

$$\begin{aligned}\nabla \cdot \mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{k}} + v_z\hat{\mathbf{k}}) \\ &= \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}, \end{aligned} \tag{11}$$

whereas

$$\begin{aligned}\mathbf{v} \cdot \nabla &= (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{k}} + v_z\hat{\mathbf{k}}) \cdot \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \\ &= v_x\frac{\partial}{\partial x} + v_y\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}. \end{aligned} \tag{12}$$

The symbol ∇, introduced by Hamilton,* is read as *del*, sometimes as *nabla* because it looks like an ancient Hebrew instrument of that name.

Closure. Given any vector field $\mathbf{v}$ (that is C^1 in the region $\mathcal{R}$ of interest), the scalar valued divergence of $\mathbf{v}$ at any point P is defined by (3), and has the physical significance of being the outflow per unit volume at P (hence the name "divergence"). Even if $\mathbf{v}$ is not a fluid velocity field (for instance, it might be an electric field, a gravitational force field, etc.) we can always *think* of it as a fluid velocity field and of div $\mathbf{v}$ as the outflow per unit volume. For computational purposes, we rely on the form (7) for Cartesian coordinates or on an analogous form for a cylindrical or spherical coordinate system. Finally, we see in (10) that the divergence of $\mathbf{v}$ amounts to a $\nabla \cdot$ operation on $\mathbf{v}$, where ∇ is the vector differential operator defined in (9); ∇ is our link to the gradient and curl in Sections 16.4 and 16.5.

EXERCISES 16.3

1. Work out div $\mathbf{v}$ for the following vector fields, each of which is defined over all of 3-space, say. Further, evaluate div $\mathbf{v}$ at $P = (3, -1, 4)$, and verify that $\mathbf{v}$ is C^1 in some region $\mathcal{R}$ containing P. (t is the time.)

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ $(a, b, c$ constants)
(b) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - y\hat{\mathbf{j}} + z^2t\hat{\mathbf{k}}$
(d) $\mathbf{v} = xyz(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 3\hat{\mathbf{k}})$

**William R. Hamilton* (1805–1865) was both a great mathematician and a great physicist. While still an undergraduate, he was named Professor of Astronomy at Trinity College and Royal Astronomer of Ireland.

(e) $\mathbf{v} = xy\hat{\mathbf{i}} - 2(x^2 + z^2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + x^2t^2\hat{\mathbf{j}}$
(h) $\mathbf{v} = \sin(x^2 + y^2 + z^2)\hat{\mathbf{j}}$

2. Determine the divergence $\operatorname{div}\mathbf{v}$ of the fluid flow field $\mathbf{v}$ given by (3) in Section 16.2.

3. Make up a vector field $\mathbf{v}(x, y, z)$ such that

(a) $\nabla \cdot \mathbf{v} = 0$ everywhere
(b) $\nabla \cdot \mathbf{v} > 0$ everywhere
(c) $\nabla \cdot \mathbf{v} < 0$ everywhere
(d) $\nabla \cdot \mathbf{v} > 0$ in $x^2 + y^2 + z^2 > 1$ and $\nabla \cdot \mathbf{v} < 0$ in $x^2 + y^2 + z^2 < 1$
(e) $\nabla \cdot \mathbf{v} > 0$ in $|x| > 1$ and $\nabla \cdot \mathbf{v} < 0$ in $|x| < 1$

4. This exercise is to promote understanding of the limit definition (3). Specifically, we ask you to evaluate $\operatorname{div}\mathbf{v}$ at $P = (0, 0, 0)$, say, by actually evaluating $(\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA)/V$ and taking the limit as $\mathcal{B} \to 0$. Take $\mathcal{B}$ to be the cube $|x| \leq \epsilon$, $|y| \leq \epsilon$, $|z| \leq \epsilon$. Show that your result agrees with that obtained (much more readily) from (7). NOTE: Do not merely mimic our steps (4)–(6). Rather, actually compute $\int \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ on each of the six faces, add the results, divide by $V = 8\epsilon^3$, and take the limit as $\epsilon \to 0$.

(a) $\mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}}$
(b) $\mathbf{v} = 3\hat{\mathbf{i}} + 4\hat{\mathbf{j}} - 2\hat{\mathbf{k}}$
(c) $\mathbf{v} = 5xe^z\hat{\mathbf{i}}$
(d) $\mathbf{v} = (x + 1)\sin y\hat{\mathbf{j}}$
(e) $\mathbf{v} = x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} - 4z^3\hat{\mathbf{k}}$
(f) $\mathbf{v} = (x^2 - 4x + yz^2)\hat{\mathbf{i}} + \hat{\mathbf{j}}$

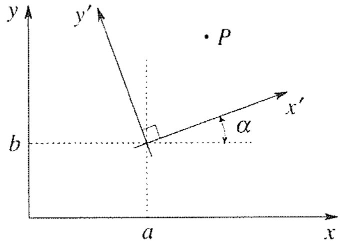
5. (*Invariance property of* $\operatorname{div}\mathbf{v}$) Let a C^1 vector field $\mathbf{v}$ be represented in terms of some particular Cartesian x, y frame as $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$. (We limit ourselves to the two-dimensional case merely to reduce the algebra; the story is essentially identical for three-dimensional fields.) Then, as derived in this section,

$$\operatorname{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}. \tag{5.1}$$

As emphasized in the text, this number will be the same, at a given point $P = (x, y)$, independent of the choice of the location and orientation of the reference coordinate system. In other words, it should be equally true that

$$\operatorname{div}\mathbf{v} = \frac{\partial v_{x'}}{\partial x'} + \frac{\partial v_{y'}}{\partial y'} \tag{5.2}$$

for any values of a, b and α (see the accompanying figure).



Here we ask you to *show* that this is true, namely, that

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = \frac{\partial v_{x'}}{\partial x'} + \frac{\partial v_{y'}}{\partial y'}. \tag{5.3}$$

HINT: First, show that

$$\begin{aligned} v_x &= (\cos\alpha)v_{x'} - (\sin\alpha)v_{y'}, \\ v_y &= (\text{etc.})v_{x'} - (\text{etc.})v_{y'}, \\ x' &= (\cos\alpha)(x - a) + (\sin\alpha)(y - b), \\ y' &= (\text{etc.})(x - a) + (\text{etc.})(y - b). \end{aligned}$$

(We leave the etceteras for you to determine.) Then use chain differentiation to express $\partial/\partial x$ and $\partial/\partial y$ as linear combinations of $\partial/\partial x'$ and $\partial/\partial y'$.

16.4 Gradient

We found in Section 16.3 that the divergence of a vector field $\mathbf{v}$ can be expressed in the form $\operatorname{div}\mathbf{v} = \nabla \cdot \mathbf{v}$, where ∇ is a vector differential operator which, in Cartesian coordinates, is given by

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}. \tag{1}$$

Thus, dotting $\boldsymbol{\nabla}$ into a vector field $\mathbf{v}$ produces a scalar field associated with $\mathbf{v}$, namely, the divergence of $\mathbf{v}$. With $\boldsymbol{\nabla}$ in hand, we wonder if it might not be profitable to consider other possible actions of $\boldsymbol{\nabla}$, specifically, $\boldsymbol{\nabla}u$ (where u is a scalar field) and $\boldsymbol{\nabla} \times \mathbf{v}$ (where $\mathbf{v}$ is a vector field). In fact, these combinations will be of great importance. Just as $\boldsymbol{\nabla} \cdot \mathbf{v}$ has a name ("the divergence of $\mathbf{v}$," or div $\mathbf{v}$, for short) which suggests its physical significance, we call $\boldsymbol{\nabla}u$ the **gradient** of u, or grad u, and we call $\boldsymbol{\nabla} \times \mathbf{v}$ the **curl** of $\mathbf{v}$, or curl $\mathbf{v}$.

The pattern is as follows:*

	Input		*Output*
$\operatorname{div}\mathbf{v} = \boldsymbol{\nabla} \cdot \mathbf{v}$:	vector field $\mathbf{v}$	→	scalar field $\boldsymbol{\nabla} \cdot \mathbf{v}$
$\operatorname{grad} u = \boldsymbol{\nabla} u$:	scalar field u	→	vector field $\boldsymbol{\nabla} u$
$\operatorname{curl}\mathbf{v} = \boldsymbol{\nabla} \times \mathbf{v}$:	vector field $\mathbf{v}$	→	vector field $\boldsymbol{\nabla} \times \mathbf{v}$

In this section we consider the gradient. In the next we turn to the curl.

As stated, we define

$$\boxed{\operatorname{grad} u \equiv \boldsymbol{\nabla} u,} \tag{2}$$

where u is a C^1 scalar field. In Cartesian coordinates,

$$\operatorname{grad} u = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) u$$

or

$$\boxed{\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}.} \tag{3}$$

EXAMPLE 1. If $u(x, y, z) = x \sin y - z^3$ and $v(x, y, z, t) = 5x^2zt$, where t is the time, then $\boldsymbol{\nabla}u = \sin y\hat{\mathbf{i}} + x\cos y\hat{\mathbf{j}} - 3z^2\hat{\mathbf{k}}$ and $\boldsymbol{\nabla}v = 10xzt\hat{\mathbf{i}} + 5x^2t\hat{\mathbf{k}}$. ▮

Directional derivative. At this point we introduce the so-called directional derivative of a scalar field $u(x, y, z)$ because it will help us to understand the gradient, and because it is important in its own right. We consider a space curve C, parametrized

*A fourth combination, $\boldsymbol{\nabla}\mathbf{v}$, is possible, but is not discussed in this book. For if we write out $\boldsymbol{\nabla}\mathbf{v}$, in Cartesian coordinates, we have (in two dimensions, for brevity)

$$\boldsymbol{\nabla}\mathbf{v} = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y}\right)(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}}) = \frac{\partial v_x}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{i}} + \frac{\partial v_y}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{j}} + \frac{\partial v_y}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{j}} + \frac{\partial v_x}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{i}}. \tag{a}$$

The objects $\hat{\mathbf{i}}\hat{\mathbf{i}}, \hat{\mathbf{j}}\hat{\mathbf{j}}, \hat{\mathbf{i}}\hat{\mathbf{j}}, \hat{\mathbf{j}}\hat{\mathbf{i}}$ are neither scalars nor vectors; they are called **dyads**, and the linear combination of dyads (a) is called a **second-order tensor**. Tensors are important in such applications as continuum mechanics, differential geometry, and the theory of relativity but are not discussed here. We refer the interested reader to the 68-page book *Tensor Analysis* by H. D. Block (Columbus, Ohio: Charles E. Merrill, 1962).

by $x = x(s)$, $y = y(s)$, and $z = z(s)$, where s is the arc length along C from some reference point on C, and we wish to compute the rate of change du/ds along C. By chain differentiation, we have

$$\frac{d}{ds}u(x(s), y(s), z(s)) = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial y}\frac{dy}{ds} + \frac{\partial u}{\partial z}\frac{dz}{ds}, \tag{4}$$

which formula holds because we have assumed $u(x, y, z)$ to be C^1. That is, chain differentiation is essentially an interpolation formula, whereby du/ds is computed as a linear combination of the rates of change $\partial u/\partial x$, $\partial u/\partial y$, and $\partial u/\partial z$ in the orthogonal coordinate directions. For such interpolation to be valid, we surely need those three partial derivatives to be continuous at the point in question (Theorem 13.4.1), and that is why we have assumed that $u(x, y, z)$ is C^1. In fact, typically, the scalar fields that arise in applications are indeed C^1, perhaps with breakdowns at one or more isolated points.

Continuing, observe that the right-hand side of (4) is a dot product:

$$\frac{du}{ds} = \left(\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}\right) \cdot \left(\frac{dx}{ds}\hat{\mathbf{i}} + \frac{dy}{ds}\hat{\mathbf{j}} + \frac{dz}{ds}\hat{\mathbf{k}}\right). \tag{5}$$

The first vector on the right is $\boldsymbol{\nabla} u$, and the second is $d\mathbf{R}/ds$, where $\mathbf{R}(s) = x(s)\hat{\mathbf{i}} + y(s)\hat{\mathbf{j}} + z(s)\hat{\mathbf{k}}$ is the position vector from the origin to the point $P = (x(s), y(s), z(s))$ on C. Not only is $d\mathbf{R}/ds$ a tangent vector to C at P, it is a *unit* vector because

$$\left\|\frac{d\mathbf{R}}{ds}\right\| = \left\|\lim_{\Delta\to 0}\frac{\Delta\mathbf{R}}{\Delta s}\right\| = \lim_{\Delta\to 0}\frac{\|\Delta\mathbf{R}\|}{\Delta s} = 1, \tag{6}$$

by the definition of arc length (Section 15.2.2). If we denote that unit tangent vector $d\mathbf{R}/ds$ as $\hat{\mathbf{s}}$, then (5) becomes

$$\boxed{\frac{du}{ds} = \boldsymbol{\nabla} u \cdot \hat{\mathbf{s}}.} \tag{7}$$

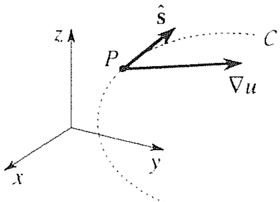


Figure 1. Computing du/ds at P.

If we wish, we can dispense with the curve C altogether. If we want the derivative du/ds, at any point P, in any desired direction defined by a unit vector $\hat{\mathbf{s}}$, then it is given by the gradient of u, at P, dotted into $\hat{\mathbf{s}}$ (Fig. 1). For instance, if $\hat{\mathbf{s}} = \hat{\mathbf{i}}$ then (7) gives $du/ds = (u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}) \cdot \hat{\mathbf{i}} = \partial u/\partial x$. Similarly, if $\hat{\mathbf{s}} = \hat{\mathbf{j}}$ then $du/ds = \partial u/\partial y$, and if $\hat{\mathbf{s}} = \hat{\mathbf{k}}$ then $du/ds = \partial u/\partial z$. More generally, however, (7) gives the rate of change of u in *any* $\hat{\mathbf{s}}$ direction.

EXAMPLE 2. Compute the directional derivative of the field $u(x, y, z) = x^2 - 3yz$ in the direction of the vector $\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}}$, at the point $(2, -1, 4)$. Using (7),

$$\frac{du}{ds}(2, -1, 4) = (2x\hat{\mathbf{i}} - 3z\hat{\mathbf{j}} - 3y\hat{\mathbf{k}})\Big|_{(2,-1,4)} \cdot \frac{\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}}}{\sqrt{6}} = -\frac{14}{\sqrt{6}}.$$

COMMENT. Remember that the gradient is a vector, but the directional derivative is a scalar. ▮

Interpretation of ∇u. We were able to interpret the scalar div $\mathbf{v}(P)$ simply, as the outflow per unit volume at P. We do not have an analogous physical interpretation of the vector ∇u, but we can use (7) to draw geometrical conclusions as to its direction and magnitude.

Consider any point P in a region throughout which a C^1 scalar field u is defined. Suppose that $\nabla u \neq \mathbf{0}$ at P and that there is a $u = \text{constant}$ surface $\mathcal{S}$ through P and a tangent plane $\mathcal{T}$ (Fig. 2); for instance, if u is a temperature field, then $\mathcal{S}$ is an isothermal surface. If $\hat{\mathbf{s}}$, at P, is chosen as any vector in the tangent plane $\mathcal{T}$, then surely du/ds must be zero. Since $du/ds = \nabla u \cdot \hat{\mathbf{s}} = 0$ for *every* $\hat{\mathbf{s}}$ at P in the tangent plane, and both ∇u and $\hat{\mathbf{s}}$ are nonzero, it follows that ∇u is normal to the tangent plane $\mathcal{T}$ and hence to the surface $\mathcal{S}$, at P.

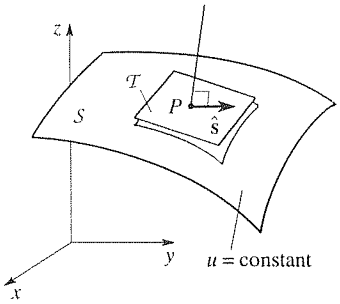


Figure 2. Geometrical interpretation of ∇u.

If, letting $\hat{\mathbf{s}}$ be in the tangent plane, we learn that ∇u is normal to $\mathcal{S}$, then to seek additional information about ∇u it seems logical to let $\hat{\mathbf{s}}$ be along the normal line at P, say in the direction of increasing u, for definiteness. Then, writing du/dn and $\hat{\mathbf{n}}$ for du/ds and $\hat{\mathbf{s}}$, respectively, (7) gives

$$\begin{aligned}\frac{du}{dn} &= \nabla u \cdot \hat{\mathbf{n}} \\ &= \|\nabla u\|\,(1)\cos 0 = \|\nabla u\|, \end{aligned} \tag{8}$$

so that the magnitude of ∇u is the directional derivative of u along the normal line to $\mathcal{S}$, in the direction of increasing u.

In summary, we can say this about the gradient ∇u of a scalar field $u(x, y, z)$ at a point P: *its direction is normal to the $u = \text{constant}$ surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.*

EXAMPLE 3. Let our scalar field be a two-dimensional temperature field $T(x, y)$, defined on the rectangle $0 \leq x \leq 2.2$, $0 \leq y \leq 1$. Suppose that T is not known analytically, but is measured experimentally, and that when the data are organized into level curves (isotherms in this case) the resulting graph is as shown in Fig. 3. The problem that we pose is to determine ∇T at $P = (1, 0.4)$. Drawing the lines PA and PB, we compute

$$\frac{\partial T}{\partial x}(P) \approx \frac{T(A) - T(P)}{\|\mathbf{PA}\|} \approx \frac{60 - 40}{0.51} \approx 39 \tag{9a}$$

and

$$\frac{\partial T}{\partial y}(P) \approx \frac{T(B) - T(P)}{\|\mathbf{PB}\|} \approx \frac{60 - 40}{0.2} \approx 100 \tag{9b}$$

so that

$$\nabla T(P) = \frac{\partial T}{\partial x}(P)\hat{\mathbf{i}} + \frac{\partial T}{\partial y}(P)\hat{\mathbf{j}} \approx 39\hat{\mathbf{i}} + 100\hat{\mathbf{j}}. \tag{10}$$

In (9a) we appealed to the definition

$$\frac{\partial T}{\partial x} = \lim_{\Delta x \to 0} \frac{T(x + \Delta x, y) - T(x, y)}{\Delta x} \tag{11}$$

but were unable to pass to the limit as $\Delta x \to 0$ because of the discrete spacing of the level curves provided in the plot. Thus, we took the smallest possible Δx, namely $\|\mathbf{PA}\|$, and "lived" with the approximation (9a). Similarly for (9b).

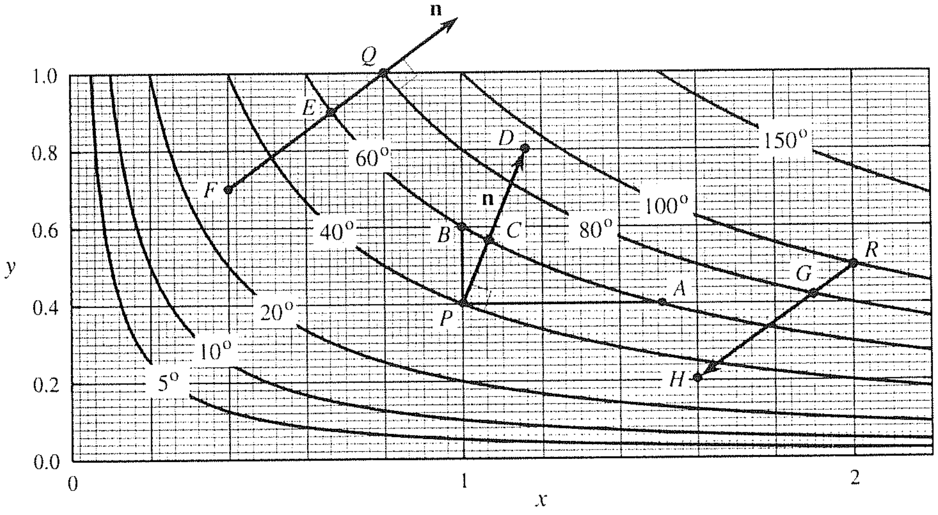


Figure 3. Level curves of $T(x, y)$.

In place of the Cartesian formula (10), we can compute ∇T from the intrinsic formula

$$\nabla T = \frac{dT}{dn}\hat{\mathbf{n}} \tag{12}$$

that is implied by the italicized summary preceding this example. Here, $\hat{\mathbf{n}}$ is the unit normal vector to the $T =$ constant curve in the direction of increasing T, and dT/dn is the directional derivative of T in the direction of $\hat{\mathbf{n}}$. Drawing the normal to the $T = 40$ curve, from P to some other convenient point, say D, we have

$$\hat{\mathbf{n}} = \frac{\mathbf{PD}}{\|\mathbf{PD}\|} \approx \frac{0.16\hat{\mathbf{i}} + 0.4\hat{\mathbf{j}}}{\sqrt{(0.16)^2 + (0.4)^2}} \approx 0.37\hat{\mathbf{i}} + 0.93\hat{\mathbf{j}} \tag{13}$$

and

$$\frac{dT}{dn} \approx \frac{T(C) - T(P)}{\|\mathbf{PC}\|} \approx \frac{60 - 40}{0.18} \approx 111 \tag{14}$$

so (12) gives

$$\nabla T \approx (111)(0.37\hat{\mathbf{i}} + 0.93\hat{\mathbf{j}}) \approx 41\hat{\mathbf{i}} + 103\hat{\mathbf{j}}. \tag{15}$$

In fact, the temperature field that we have plotted in Fig. 3 is actually $T(x, y) = 100xy$. Thus, the exact value of ∇T at $(1, 0.4)$ is

$$\nabla T = \frac{\partial T}{\partial x}\hat{\mathbf{i}} + \frac{\partial T}{\partial y}\hat{\mathbf{j}} = 100y\hat{\mathbf{i}} + 100x\hat{\mathbf{j}} = 40\hat{\mathbf{i}} + 100\hat{\mathbf{j}}, \tag{16}$$

in view of which our estimates (10) and (15) are seen to be respectable, considering that our Δx, Δy, and Δn increments were not very small. More sophisticated difference quotients could have been used, to squeeze out more accuracy, but the thrust of this example is to illustrate concepts, not to maximize numerical accuracy.

As a second calculation, using Fig. 3, let us compute ∇T at $Q = (0.8, 1)$. Let us use (12). In this case we cannot use a forward difference quotient for dT/dn because Q is at a boundary point of the region and a forward step would carry us outside the region, where there is no data.* Thus, use a backward difference quotient instead. First, draw a normal line FQ. Then,

$$\hat{\mathbf{n}} = \frac{\mathbf{FQ}}{\|\mathbf{FQ}\|} \approx \frac{0.4\hat{\mathbf{i}} + 0.3\hat{\mathbf{j}}}{\sqrt{(0.4)^2 + (0.3)^2}} \approx 0.8\hat{\mathbf{i}} + 0.6\hat{\mathbf{j}}, \tag{17}$$

$$\frac{dT}{dn} \approx \frac{T(Q) - T(F)}{\|\mathbf{QF}\|} \approx \frac{80 - 60}{0.17} \approx 118 \tag{18}$$

so

$$\nabla T(0.8, 1) \approx (118)(0.8\hat{\mathbf{i}} + 0.6\hat{\mathbf{j}}) \approx 94\hat{\mathbf{i}} + 71\hat{\mathbf{j}} \tag{19}$$

compared with the exact value $100\hat{\mathbf{i}} + 80\hat{\mathbf{j}}$.

Finally, let us compute dT/ds at $R = (2, 0.5)$, in the direction $\mathbf{RH}$, where H is at $(1.6, 0.2)$. We have

$$\frac{dT}{ds}(R) \approx \frac{T(G) - T(R)}{\|\mathbf{GR}\|} \approx \frac{80 - 100}{0.13} \approx -154 \tag{20}$$

compared with the exact value -160. ∎

Closure. Having met the vector differential operator ∇, in Section 16.3, we define the gradient of a scalar field u as $\operatorname{grad} u = \nabla u$. In Cartesian coordinates,

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}. \tag{21}$$

Next, we derived the formula

$$\frac{du}{ds} = \nabla u \cdot \hat{\mathbf{s}} \tag{22}$$

for the directional derivative du/ds in any given $\hat{\mathbf{s}}$ direction. Besides being of importance in its own right, (22) help us to understand the gradient ∇u. We learn

*In the difference quotient definition of the derivative,

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}, \tag{a}$$

Δx can tend to zero through positive values, in which case we call the difference quotient in (a) a **forward difference quotient**, or through negative values. In the latter case we can re-express (a) as

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x) - f(x - \Delta x)}{\Delta x}, \tag{b}$$

and call it a **backward difference quotient**.

that ∇u at any point P is normal to the u = constant surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.

EXERCISES 16.4

1. Work out grad u for the following scalar fields, each of which is defined over all of 3-space. Further, evaluate grad u at $(9, 4, -1)$.

(a) $u = 6x$ (b) $u = x^2$
(c) $u = z \sin(x^2 + y^2)$ (d) $u = x^3$
(e) $u = xyz$ (f) $u = x^2 + y^2 + z^2$
(g) $u = x^2 + z^2$ (h) $u = x - y$
(i) $u = x + y + z$ (j) $u = z - x^2 - y^2$

2. Evaluate the directional derivative du/ds, at the designated point P, in the direction of the given vector $\mathbf{v}$.

(a) $u = x^2 + y^2 + z^2, \ P = (2, 1, 5), \ \mathbf{v} = \hat{\mathbf{i}}$
(b) $u = x + y + 3z, \ P = (1, 0, 3), \ \mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 5\hat{\mathbf{k}}$
(c) $u = xyz, \ P = (1, -1, 2), \ \mathbf{v} = 3\hat{\mathbf{i}} - \hat{\mathbf{k}}$
(d) $u = xyz, \ P = (3, 2, 1), \ \mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$
(e) $u = \sqrt{x(y + z)}, \ P = (1, 2, 3), \ \mathbf{v} = \hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}$
(f) $u = 3x^2, \ P = (2, 6, -1), \ \mathbf{v} = 2\hat{\mathbf{j}} + 5\hat{\mathbf{k}}$

3. For the temperature field $T(x, y)$ given in Fig. 3, evaluate the following quantities directly from the figure, with the help of suitable difference quotients, and compare your results with the exact values obtained from the expression $T = 100xy$.

(a) ∇T at $(0.4, 1)$ (b) ∇T at $(1, 1)$
(c) ∇T at $(1, 0.2)$ (d) ∇T at $(0.5, 0.2)$
(e) ∇T at $(1.6, 0.5)$ (f) ∇T at $(1.5, 1)$
(g) dT/ds at $(0.4, 1)$ in the direction of $-\hat{\mathbf{i}} - \hat{\mathbf{j}}$
(h) dT/ds at $(1, 1)$ in the direction of $-\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$
(i) dT/ds at $(1, 0.2)$ in the direction of $\hat{\mathbf{i}} + \hat{\mathbf{j}}$
(j) dT/ds at $(0.5, 0.2)$ in the direction of $-\hat{\mathbf{j}}$
(k) dT/ds at $(1.6, 0.5)$ in the direction of $\hat{\mathbf{i}} - \hat{\mathbf{j}}$
(l) dT/ds at $(1.5, 1)$ in the direction of $-\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$

4. Let the electric potential (i.e., the voltage) be given by $V(x, y, z) = 3x^2y - xz$. If a positive charge is placed at $P = (x, y, z)$, in what direction will the charge begin to move? NOTE: It is known, from electric field theory, that such a charge will begin to move in the direction of maximum rate of voltage drop.

(a) $P = (2, 3, -1)$ (b) $P = (4, 0, -1)$
(c) $P = (1, 2, 5)$ (d) $P = (0, 0, 4)$
(e) $P = (0, 2, 1)$ (f) $P = (3, -2, 0)$

5. (*Convective derivative*) Let $\mathbf{v}(x, y, z, t) = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$ be a fluid velocity field, and consider some scalar property of the flow, such as the temperature field $T(x, y, z, t)$. If we swim along any desired path according to $x = x(t)$, $y = y(t)$, $z = z(t)$, then the timewise rate of change of T that we observe is, by chain differentiation,

$$\frac{d}{dt} T(x(t), y(t), z(t)) = \frac{\partial T}{\partial t} + \frac{\partial T}{\partial x}\frac{dx}{dt} + \frac{\partial T}{\partial y}\frac{dy}{dt} + \frac{\partial T}{\partial z}\frac{dz}{dt}. \tag{5.1}$$

(a) If we choose to *drift* with the fluid, then $dx/dt = v_x$, $dy/dt = v_y$, and $dz/dt = v_z$. In this case, show that (a) becomes

$$\frac{dT}{dt} = \frac{\partial T}{\partial t} + (\mathbf{v} \cdot \nabla)T. \tag{5.2}$$

This is often called the **convective derivative** because it is the derivative obtained when we drift, or convect, with the fluid. The special notation D/Dt, suggested by *Sir G. G. Stokes* (1819–1903), is often used:

$$\boxed{\frac{D(\ )}{Dt} = \frac{\partial(\ )}{\partial t} + (\mathbf{v} \cdot \nabla)(\ ).} \tag{5.3}$$

(b) To understand the nature of the contributions of each of the two terms on the right-hand side of (5.2) apply (5.2) to the three simple cases: $T = 2t$, $\mathbf{v} = \mathbf{0}$; $T = 3x$, $\mathbf{v} = U\hat{\mathbf{i}}$; $T = 2t + 3x$, $\mathbf{v} = U\hat{\mathbf{i}}$. Include whatever words of explanation (and sketches) seem appropriate.

(c) Does it matter if we rewrite $\mathbf{v} \cdot \nabla$ in equation (5.3) as $\nabla \cdot \mathbf{v}$? That is, is it true that

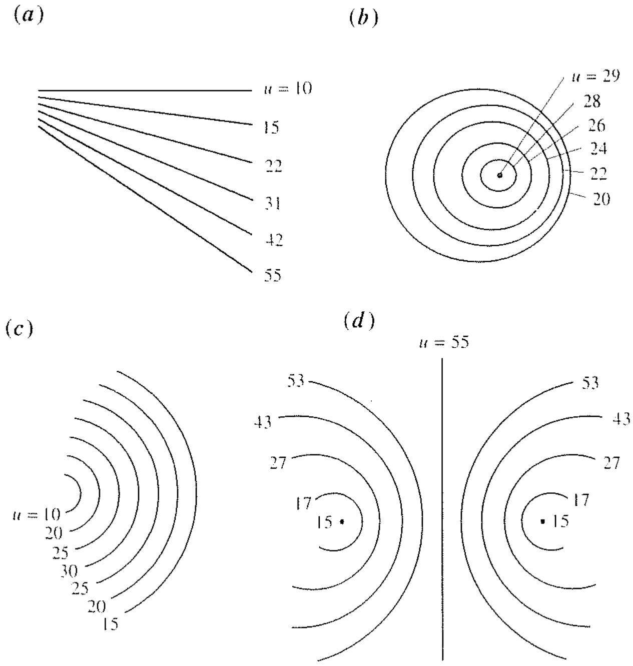
$$\frac{D(\ )}{Dt} = \frac{\partial(\ )}{\partial t} + (\operatorname{div} \mathbf{v})(\ )?$$

6. Evaluate the rate of change of temperature, dT/dt, measured by a thermometer that drifts along with the fluid velocity $\mathbf{v}$. Evaluate dT/dt at $x = 2$, $y = -1$, $z = 3$, and $t = 4$. HINT: Read Exercise 5.

(a) $T = 100t^2 + 2(x^2 + y^2),\ \mathbf{v} = 20xt\hat{\mathbf{i}} - 10y\hat{\mathbf{j}}$
(b) $T = 25xt^2 - 2xyz,\ \mathbf{v} = 20\hat{\mathbf{i}} + xy\hat{\mathbf{j}} - t\hat{\mathbf{k}}$
(c) $T = (x + y + z)t,\ \mathbf{v} = xz\hat{\mathbf{i}} - xt\hat{\mathbf{j}}$
(d) $T = y^2 - zt,\ \mathbf{v} = \hat{\mathbf{i}} + 2x^2t\hat{\mathbf{j}} - \hat{\mathbf{k}}$

7. Define $u(x, y)$ to be 1 along the curve $y = x^2$ except for the origin $x = y = 0$, and define it to be zero at all other points in the plane, including the origin. Show that du/ds exists and equals 0 at the origin, for *all* directions $\hat{\mathbf{s}}$, but that $u(x, y)$ is, nevertheless, discontinuous at the origin. NOTE: Recall from the calculus that if $y'(x)$ exists at x_0, then $y(x)$ is necessarily continuous at x_0. For functions of more than one variable, we see, from this example, that the existence of directional derivatives in *every direction*, at a given point, does *not* suffice to imply that the function is continuous at that point.

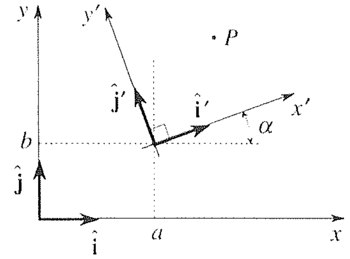
8. Some level curves of a scalar field $u(x, y)$ are sketched below. Trace or photocopy that sketch, and indicate the point(s), if any, at which $\nabla u \approx \mathbf{0}$, as well as the point(s) at which ∇u takes on its greatest magnitude.



9. (*Invariance property of* ∇u) In Section 16.3 we stress the invariance of the divergence of $\mathbf{v}$; that is, the value of $\nabla \cdot \mathbf{v}$ at a given field point is the same, whether we use one Cartesian coordinate system or another. The same invariance holds for the gradient. Prove that claim – in two dimensions, for simplicity. That is, show that

$$\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} = \frac{\partial u}{\partial x'}\hat{\mathbf{i}}' + \frac{\partial u}{\partial y'}\hat{\mathbf{j}}', \tag{9.1}$$

where the x, y and x', y' systems differ by an arbitrary displacement and an arbitrary rotation as shown in the figure.



10. (*Limit definition of gradient*) recall that we began with the limit definition of the divergence,

$$\operatorname{div} \mathbf{v}(P) \equiv \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\}. \tag{10.1}$$

Working out the right-hand side, for the case of Cartesian coordinates, gave

$$\operatorname{div} \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z},$$

and expressing the latter as

$$\begin{aligned} \operatorname{div} \mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z} \right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \\ &= \nabla \cdot \mathbf{v} \end{aligned}$$

led us to the del operator,

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}.$$

With the ∇ operator in hand, we then used ∇ to introduce the gradient, as $\operatorname{grad} u = \nabla u$. Alternatively, the gradient can be introduced intrinsically, by a limit definition analogous to the definition (10.1) of the divergence, namely,

$$\boxed{\operatorname{grad} u(P) \equiv \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V} \right\}.} \tag{10.2}$$

Following steps analogous to those in Section 16.3, derive from (10.2) the Cartesian coordinate expression of $\operatorname{grad} u$ as

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}$$

which, of course, is the same as ∇u.

16.5 Curl

Having considered the divergence and gradient, $\operatorname{div}\mathbf{v} = \nabla\cdot\mathbf{v}$ and $\operatorname{grad} u = \nabla u$, we now turn to the **curl of v**, defined here as

$$\boxed{\operatorname{curl}\mathbf{v} \equiv \nabla\times\mathbf{v},} \tag{1}$$

where $\mathbf{v}$ is a C^1 vector field. Working out the right-hand side of (1) in Cartesian coordinates yields

$$\begin{aligned}\operatorname{curl}\mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right)\times\left(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}\right)\\ &= \begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z}\\ v_x & v_y & v_z\end{vmatrix}\end{aligned} \tag{2}$$

or*

$$\boxed{\operatorname{curl}\mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}.} \tag{3}$$

EXAMPLE 1. If $\mathbf{v}(x,y,z,t) = xyz\hat{\mathbf{i}} - 2y^2t\hat{\mathbf{k}}$, then

$$\begin{aligned}\operatorname{curl}\mathbf{v} &= (-4yt - 0)\hat{\mathbf{i}} - (0 - xy)\hat{\mathbf{j}} + (0 - xz)\hat{\mathbf{k}}\\ &= -4yt\hat{\mathbf{i}} + xy\hat{\mathbf{j}} - xz\hat{\mathbf{k}}. \quad ▮\end{aligned}$$

To interpret curl v physically, let $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$ be a fluid velocity field. Let us examine curl v which, in the study of fluid mechanics, is called the **vorticity**

*Evidently, if we expand the determinant about the *first row*, we do obtain the terms on the right-hand side of (3). Yet if we expand about the third row, say, we obtain

$$v_x\left(\hat{\mathbf{j}}\frac{\partial}{\partial z} - \frac{\partial}{\partial y}\hat{\mathbf{k}}\right) - v_y\left(\hat{\mathbf{i}}\frac{\partial}{\partial z} - \frac{\partial}{\partial x}\hat{\mathbf{k}}\right) + v_z\left(\hat{\mathbf{i}}\frac{\partial}{\partial y} - \frac{\partial}{\partial x}\hat{\mathbf{j}}\right), \tag{a}$$

which is not the same. The point to bear in mind is that ∇ is located to the left of $\mathbf{v}$ in (1) so that $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ act on v_x, v_y, and v_z. Thus, the terms in (a) need to be rearranged: the first term, $v_x\hat{\mathbf{j}}\,\partial/\partial z$, should be rearranged as $(\partial v_x/\partial z)\hat{\mathbf{j}}$, and so on, for each of the other terms. If we remember to do that, the correct expression is obtained no matter which row or column is expanded about.

vector. For simplicity, let $\mathbf{v}$ be a *plane flow*, where $v_z = 0$ and v_x, v_y do not vary with z. Thus, $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$ and (3) reduces to

$$\operatorname{curl}\mathbf{v} = \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}. \tag{4}$$

Consider, heuristically, the motion of the little rectangular element of fluid shown in Fig. 1. At time t it is in position 1. If the element were *rigid*, its motion would consist of a translation plus a rotation. If so, then at time $t + \Delta t$ it might be in configuration 2, say, due to a translation OO' plus a rotation $\Delta\alpha$, and its angular velocity (taken as positive counterclockwise, with the right-hand rule used to assign a vector direction) would be $\boldsymbol{\omega} \sim (\Delta\alpha/\Delta t)\hat{\mathbf{k}}$. However, a fluid element is *deformable*, not rigid, so it may also suffer a shear deformation as indicated in the configuration 3. What, then, is the fluid's "angular velocity" to mean? It is *defined*, in fluid mechanics, as the average angular velocity of initially perpendicular edges such as OA and OB. Then

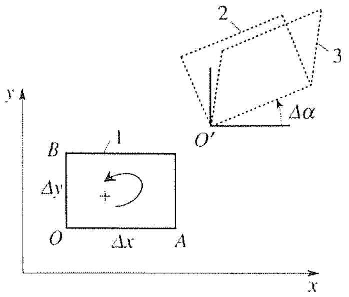


Figure 1. Plane motion of fluid element.

$$\text{angular velocity of } OA = \frac{v_y(A) - v_y(O)}{\Delta x}\hat{\mathbf{k}} \to \frac{\partial v_y}{\partial x}\hat{\mathbf{k}} \quad \text{as } \Delta x \to 0, \tag{5a}$$

$$\text{angular velocity of } OB = \frac{v_x(O) - v_x(B)}{\Delta y}\hat{\mathbf{k}} \to -\frac{\partial v_x}{\partial y}\hat{\mathbf{k}} \quad \text{as } \Delta y \to 0, \tag{5b}$$

so that the

$$\text{average of the two} \equiv \boldsymbol{\omega} = \frac{1}{2}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}. \tag{6}$$

In case (5a) is not clear, consider Fig. 2. The average of the x velocity components $v_x(O)$ and $v_x(A)$ produce the x-wise translation of OA, their difference produces the stretching or contracting of OA, and the average of $v_y(O)$ and $v_y(A)$ produce the y-wise translation of OA. But the *angular velocity* of OA is due to the difference of $v(O)$ and $v(A)$. It is the vertical velocity of A relative to O, namely, $v_y(A) - v_y(O)$ divided by the radius OA of that motion. Similarly, the angular velocity of OB is $v_x(O) - v_x(B)$ divided by OB.

Comparing (4) and (6), we see that $\operatorname{curl}\mathbf{v} = 2\boldsymbol{\omega}$, from which we draw the following simple and important physical interpretation of the curl of a vector field: $\operatorname{curl}\mathbf{v}(P)$ *is twice the angular velocity of the fluid at* P.

Although our (heuristic) proof of the italicized claim was for plane flows, it is not hard to show that that result holds for nonplanar flows as well. And if our vector field is not a fluid velocity field (e.g., it might be an electric field, magnetic field, or gravitational force field) we can at least think of it as a velocity field in order to have access to the physical interpretation stated above in italics.

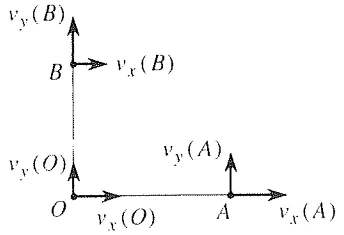


Figure 2. Angular velocities of OA and OB.

EXAMPLE 2. *Shear Flow.* To illustrate, consider the simple flow

$$\mathbf{v} = \kappa y\hat{\mathbf{i}}, \tag{7}$$

where κ is a positive constant. Shown in Fig. 3,the flow is known as a *shear flow* and is similar to the flow in the *boundary layer* on an aircraft wing. Considering a typical

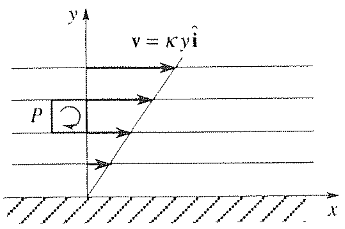


Figure 3. Shear flow.

fluid particle such as P, observe that the top of the particle is translating faster than the bottom so the particle undergoes a clockwise angular velocity as it translates to the right. Furthermore, it appears that all particles have the same angular velocity (though different translational velocities) since, for a given particle size, the velocity differential from top to bottom of the particle is the same for each particle. Thus, since curl v is twice the angular velocity of the fluid particle, and the latter is evidently a constant and in the $-\hat{k}$ direction (by the right-hand rule), we expect curl v to be a constant and in the $-\hat{k}$ direction. Let us see:

$$\operatorname{curl} \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ \kappa y & 0 & 0 \end{vmatrix} = -\kappa\hat{\mathbf{k}}, \tag{8}$$

which is a constant and in the $-\hat{k}$ direction, as expected. ▮

It is interesting to observe that, in itself, the streamline pattern tells us nothing about the vorticity, curl v. For instance, the three flows $\mathbf{v} = 3y\hat{\mathbf{i}}$, $-3y\hat{\mathbf{i}}$, and $3\hat{\mathbf{i}}$ all have the same streamline pattern: the set of horizontal lines, as in Fig. 3. But in these cases the fluid particles are rotating clockwise, counterclockwise, and not at all, respectively. To carry this point a bit further, consider the flow over a semicircular bump, shown in Fig. 4 of Section 16.2. It seems reasonable to imagine the fluid particles moving and rotating as sketched here in Fig. 4. Yet,

$$\mathbf{v} = U\hat{\mathbf{i}} + \frac{Ua^2}{(x^2+y^2)^2}\left[(y^2-x^2)\hat{\mathbf{i}} - 2xy\hat{\mathbf{j}}\right]$$

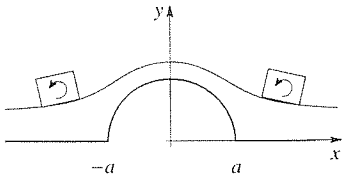


Figure 4. Particles rotating like this?

and, working out curl v, we find that terms cancel so that curl v = 0 everywhere in the field. Thus, the particles are not rotating at all even as they move along curved streamlines, in the same way that seats on a Ferris wheel remain horizontal even as they move in a vertical circle.

EXAMPLE 3. *Maxwell's Equations.* To illustrate the role of the curl in applications, we note that the equations

$$\nabla \times \mathbf{E} = \mathbf{0}, \tag{9a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}, \tag{9b}$$

are two of the four classical **Maxwell's equations** governing steady (i.e., not varying with time) electromagnetic fields. **E**, **H**, and **J** are the electric field intensity, magnetic field intensity, and the current density, respectively. ▮

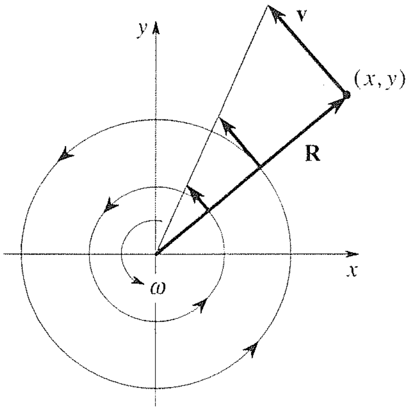
Closure. We define the curl of a vector field v as $\nabla \times \mathbf{v}$. In Cartesian coordinates the latter can be evaluated by (3), but it is easier to remember the determinant form given in (2). To attach a physical significance to $\nabla \times \mathbf{v}$, think of v as a fluid velocity field (whether it is or not). Then $\nabla \times \mathbf{v}$ at any given point P is twice the angular velocity of the fluid at P.

EXERCISES 16.5

1. Work out curl $\mathbf{v}$ for the following vector fields. Further, evaluate curl $\mathbf{v}$ at $(3, 4, -1)$.

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ (a, b, c constants)
(b) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - y^2\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
(d) $\mathbf{v} = xyz(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 3\hat{\mathbf{k}})$
(e) $\mathbf{v} = xy\hat{\mathbf{i}} - 2(x^2 + z^2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + x^2\hat{\mathbf{j}}$
(h) $\mathbf{v} = \sin(x + y + z)\hat{\mathbf{j}}$
(i) $\mathbf{v} = f(x)\hat{\mathbf{i}} + g(y)\hat{\mathbf{j}} + h(z)\hat{\mathbf{k}}$
(j) $\mathbf{v} = f(y)\hat{\mathbf{i}} + g(x)\hat{\mathbf{j}}$

2. (*Solid-body rotation*) Consider a plane fluid flow that is a counterclockwise "solid-body rotation" about the z axis with angular velocity $\boldsymbol{\omega} = \omega\hat{\mathbf{k}}$, as sketched in the figure. That is, the fluid might just as well be frozen solid, and spun about the z axis. Introducing polar coordinates r, θ, the position vector is



$$\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} = r\cos\theta\hat{\mathbf{i}} + r\sin\theta\hat{\mathbf{j}}. \tag{2.1}$$

Since $\dot{r} = 0$ and $\dot{\theta} = \omega$ (where dots denote time derivatives), we have for the fluid velocity field

$$\begin{aligned} \mathbf{v} = \dot{\mathbf{R}} &= r(-\dot{\theta}\sin\theta\hat{\mathbf{i}} + \dot{\theta}\cos\theta\hat{\mathbf{j}}) \\ &= \omega(-r\sin\theta\hat{\mathbf{i}} + r\cos\theta\hat{\mathbf{j}}) = \omega(-y\hat{\mathbf{i}} + x\hat{\mathbf{j}}). \end{aligned} \tag{2.2}$$

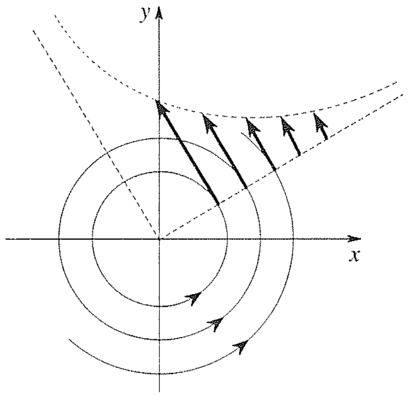
(a) Work out curl $\mathbf{v}$ for the $\mathbf{v}$ field given by (2.2) and explain, in physical terms, why the result makes sense.

(b) Derive (2.2) differently, by differentiating $\mathbf{R} = r\hat{\mathbf{e}}_r$ with respect to time.

3. (*Vortex flow*) The fluid velocity field

$$\mathbf{v} = \frac{\Gamma}{2\pi r}\hat{\mathbf{e}}_\theta \qquad (\Gamma = \text{constant}) \tag{3.1}$$

is purely tangential, i.e., there is no radial ($\hat{\mathbf{e}}_r$) component so that the streamlines are concentric circles centered at the origin, as sketched in the accompanying figure. Furthermore, the magnitude of the tangential velocity tends to zero as $r \to \infty$, and to ∞ as $r \to 0$. [Equation (3.1) is said to give the flow induced by a *vortex* of strength Γ located at the origin.] The problem that we pose is to re-express (3.1) in terms of $x, y, \hat{\mathbf{i}}, \hat{\mathbf{j}}$ and show that $\nabla \times \mathbf{v} = \mathbf{0}$ everywhere in the field (except at the origin, where $\nabla \times \mathbf{v}$ is not defined) so that every fluid particle (not at the origin) has no spin. [This result may seem strange in view of the fact that the particles are in circular motion! Evidently, the particles must maintain their spatial orientation (much as the seats on a Ferris wheel do) as they carry out their circular motion.]



4. (*Invariance property of curl* $\mathbf{v}$) Let a C^1 vector field $\mathbf{v}$ be represented in terms of some particular x, y frame as $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$. (We limit ourselves to the two-dimensional case merely to reduce the algebra.) Then, according to (3),

$$\text{curl}\,\mathbf{v} = \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}. \tag{4.1}$$

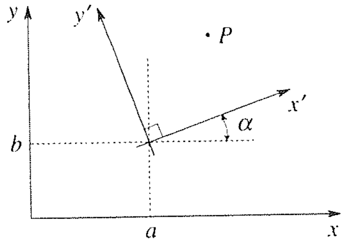
Show that this vector is the same even if a different coordinate system is used, say, the x', y' system shown in the figure. That is, show that

$$\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}} = \left(\frac{\partial v_{y'}}{\partial x'} - \frac{\partial v_{x'}}{\partial y'}\right)\hat{\mathbf{k}}'$$

or, since $\hat{\mathbf{k}}' = \hat{\mathbf{k}}$, that

$$\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} = \frac{\partial v_{y'}}{\partial x'} - \frac{\partial v_{x'}}{\partial y'} \tag{4.2}$$

for any values of a, b and α.



5. Show whether or not $\nabla \times \mathbf{v}$ is necessarily orthogonal to $\mathbf{v}$. HINT: That is, is $\mathbf{v} \cdot \nabla \times \mathbf{v}$ necessarily zero? Write

$$\mathbf{v} \cdot \nabla \times \mathbf{v} = \begin{vmatrix} v_x & v_y & v_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_x & v_y & v_z \end{vmatrix},$$

and see if the determinant is necessarily zero. It is true that two rows are identical, which seems to imply that the determinant must be zero, but be careful, and recall the footnote associated with equation (3).

6. (*Limit definition of curl*) First, read Exercise 10 in Section 16.4. Similarly, we can introduce the curl by the limit definition

$$\boxed{\operatorname{curl} \mathbf{v}(P) \equiv \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V} \right\},} \tag{6.1}$$

and deduce from (6.1) that curl $\mathbf{v}$ is indeed $\nabla \times \mathbf{v}$. That is what we ask you to do: using Cartesian coordinates, derive from (6.1) the result

$$\begin{aligned} \operatorname{curl} \mathbf{v} = {} & \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} \\ & + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}, \end{aligned} \tag{6.2}$$

which, as claimed above, is the same as $\nabla \times \mathbf{v}$. NOTE: Let us write the three results together, so the pattern among them can be seen:

$$\operatorname{div} \mathbf{v} = \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\} = \nabla \cdot \mathbf{v}, \tag{6.3a}$$

$$\operatorname{grad} u = \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V} \right\} = \nabla u, \tag{6.3b}$$

and

$$\operatorname{curl} \mathbf{v} = \lim_{B \to 0} \left\{ \frac{\int_S \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V} \right\}. \tag{6.3c}$$

16.6 Combinations; Laplacian

We have introduced the div, grad, and curl operators, and found that all three can be expressed in terms of ∇. Specifically, div $= \nabla \cdot$, grad $= \nabla$, and curl $= \nabla \times$.

In this section, we consider combinations: one of the operators above acting on a combination of fields such as $\nabla(\alpha u + \beta v)$ and $\nabla \cdot (u\mathbf{v})$, or a combination of the operators above acting on a single field, such as $\nabla \cdot \nabla u$ and $\nabla \times (\nabla u)$. Beginning with the former, the action of one such operator on a combination of fields, we note the following results:

If α, β are scalars, then

$$\nabla \cdot (\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha \nabla \cdot \mathbf{u} + \beta \nabla \cdot \mathbf{v}, \tag{1}$$

$$\nabla(\alpha u + \beta v) = \alpha\nabla u + \beta\nabla v, \tag{2}$$
$$\nabla \times (\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\nabla \times \mathbf{u} + \beta\nabla \times \mathbf{v} \tag{3}$$

so that div, grad, and curl are *linear* operators. Furthermore,

$$\nabla \cdot (u\mathbf{v}) = \nabla u \cdot \mathbf{v} + u\nabla \cdot \mathbf{v}, \tag{4}$$
$$\nabla \times (u\mathbf{v}) = \nabla u \times \mathbf{v} + u\nabla \times \mathbf{v}, \tag{5}$$
$$\nabla \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \nabla \times \mathbf{u} - \mathbf{u} \cdot \nabla \times \mathbf{v}, \tag{6}$$
$$\nabla \times (\mathbf{u} \times \mathbf{v}) = \mathbf{u}\nabla \cdot \mathbf{v} - \mathbf{v}\nabla \cdot \mathbf{u} + (\mathbf{v} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\mathbf{v}, \tag{7}$$
$$\nabla(\mathbf{u} \cdot \mathbf{v}) = (\mathbf{u} \cdot \nabla)\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{u} + \mathbf{u} \times (\nabla \times \mathbf{v}) + \mathbf{v} \times (\nabla \times \mathbf{u}), \tag{8}$$

where $(\mathbf{u} \cdot \nabla)\mathbf{v}$ in (7) and (8) means

$$\begin{aligned}(\mathbf{u} \cdot \nabla)\mathbf{v} &= (\mathbf{u} \cdot \nabla)(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \\ &= (\mathbf{u} \cdot \nabla v_x)\hat{\mathbf{i}} + (\mathbf{u} \cdot \nabla v_y)\hat{\mathbf{j}} + (\mathbf{u} \cdot \nabla v_z)\hat{\mathbf{k}} \\ &= \left(u_x \frac{\partial v_x}{\partial x} + u_y \frac{\partial v_x}{\partial y} + u_z \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{i}} + (\text{etc.})\hat{\mathbf{j}} + (\text{etc.})\hat{\mathbf{k}},\end{aligned}$$

and similarly for $(\mathbf{v} \cdot \nabla)\mathbf{u}$.

Derivation of these results is simple, and we will limit our discussion to the verification of one representative case, (4):

$$\begin{aligned}\nabla \cdot (u\mathbf{v}) &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (uv_x\hat{\mathbf{i}} + uv_y\hat{\mathbf{j}} + uv_z\hat{\mathbf{k}}) \\ &= \frac{\partial}{\partial x}(uv_x) + \frac{\partial}{\partial y}(uv_y) + \frac{\partial}{\partial z}(uv_z) \\ &= \frac{\partial u}{\partial x}v_x + \frac{\partial u}{\partial y}v_y + \frac{\partial u}{\partial z}v_z + u\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) \\ &= \nabla u \cdot \mathbf{v} + u\nabla \cdot \mathbf{v}.\end{aligned}$$

The form of this result is not surprising and is reminiscent of the (scalar) statement $(d/dx)(uv) = (du/dx)v + u(dv/dx)$ from the calculus. That is, ∇ is a differential operator, and $u\mathbf{v}$ is a product so we get two terms: one in which ∇ acts on u, and one in which it acts on $\mathbf{v}$.

Next, look at the action of a combination of div, grad, and curl operators on a single field. One such case that will be of great importance to us is the operator $\nabla \cdot \nabla$, that is, div grad. This operator is called the **Laplacian** operator,* or just the Laplacian, and is denoted by the symbol ∇^2. That is,

$$\nabla^2 \equiv \nabla \cdot \nabla = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right)$$

*After *Pierre Simon de Laplace* (1749–1827). Some authors use the symbol Δ in place of ∇^2.

or,

$$\boxed{\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.} \tag{9}$$

Read as *del square*, ∇^2 does *not* mean $\boldsymbol{\nabla}$ times $\boldsymbol{\nabla}$. It is a mathematical symbol that means $\boldsymbol{\nabla}$ *dot* $\boldsymbol{\nabla}$, in the same way that the matrix symbol $\mathbf{A}^{-1}$ means the inverse matrix of $\mathbf{A}$, not $1/\mathbf{A}$. The Laplacian can act on scalar fields *or* on vector fields:

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}, \tag{10}$$

$$\begin{aligned}\nabla^2 \mathbf{v} &= \nabla^2(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}})\\ &= \hat{\mathbf{i}}\nabla^2 v_x + \hat{\mathbf{j}}\nabla^2 v_y + \hat{\mathbf{k}}\nabla^2 v_z\\ &= \left(\frac{\partial^2 v_x}{\partial x^2} + \frac{\partial^2 v_x}{\partial y^2} + \frac{\partial^2 v_x}{\partial z^2}\right)\hat{\mathbf{i}} + (\text{etc.})\hat{\mathbf{j}} + (\text{etc.})\hat{\mathbf{k}}.\end{aligned} \tag{11}$$

Besides the div grad combination, just mentioned, some others to be noted are

$$\boxed{\begin{aligned}\text{div curl}\,\mathbf{v} &= \boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\times\mathbf{v} = 0,\\ \text{curl grad}\,u &= \boldsymbol{\nabla}\times\boldsymbol{\nabla}u = \mathbf{0},\end{aligned}} \tag{12,13}$$

$$\text{curl curl}\,\mathbf{v} = \text{curl}^2\mathbf{v} = \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{v}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\mathbf{v}) - \nabla^2\mathbf{v}, \tag{14}$$

which hold if $\mathbf{v}$ and u are C^2.

We have drawn attention to (12) and (13), by placing them in a box because they will be important to us later, and also because they are striking results, striking in the sense that the divergence of the curl of *every* C^2 vector field is found to be zero, and the curl of the gradient of *every* C^2 scalar field is found to be zero.

As a mnemonic device to remember (12) and (13), think of the letters d, g, and c (for divergence, gradient, and curl). The only combinations that come readily to mind (for the author, at least) are **dc** (for direct current), and **cg** (for center of gravity). This is, the divergence of a curl is zero, and the curl of a gradient is zero.

EXAMPLE 1. If $u = x^2y$ and $\mathbf{v} = x^2y\hat{\mathbf{i}} - z^3\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$, then

$$\begin{aligned}\nabla^2 u &= 2y \qquad \text{by}(10),\\ \nabla^2\mathbf{v} &= (2y+0+0)\hat{\mathbf{i}} + (0+0-6z)\hat{\mathbf{j}} + (0+0+0)\hat{\mathbf{k}} \qquad \text{by}(11)\\ &= 2y\hat{\mathbf{i}} - 6z\hat{\mathbf{j}},\\ \boldsymbol{\nabla}\cdot\boldsymbol{\nabla}\times\mathbf{v} &= 0 \qquad \text{by}(12),\\ \boldsymbol{\nabla}\times\boldsymbol{\nabla}u &= \mathbf{0} \qquad \text{by}(13),\\ \boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\mathbf{v}) &= \boldsymbol{\nabla}(2xy) - (2y\hat{\mathbf{i}} - 6z\hat{\mathbf{j}}) \qquad \text{by}(14),\\ &= (2x+6z)\hat{\mathbf{j}}.\ \blacksquare\end{aligned}$$

Closure. The identities (1)–(8) and (12)–(14) are of general utility in vector analysis somewhat as the various trigonometric identities are of utility in the calculus, and we will use most of them in the remainder of this chapter. The Laplace operator ∇^2 will be central in our study of partial differential equations in Chapters 18–20.

EXERCISES 16.6

1. Evaluate $\nabla^2 u$ and $\nabla \times \nabla u$ in each case.

(a) $u = x$
(b) $u = x^2 + y^2$
(c) $u = x^2 + y^2 + z^2$
(d) $u = x^2 y^3 - 6$
(e) $u = xe^y$
(f) $u = \cos(x - 2y)$
(g) $u = xyz$
(h) $u = \sin(xyz)$
(i) $u = ax + by + cz + d \quad (a, b, c, d$ constants$)$

2. Evaluate $\nabla \cdot \nabla \times \mathbf{v}$ and $\nabla \times (\nabla \times \mathbf{v})$ in each case.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$ (b) $\mathbf{v} = xe^y\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(c) $\mathbf{v} = z^3\hat{\mathbf{i}}$ (d) $\mathbf{v} = yz\hat{\mathbf{i}} - x^2\hat{\mathbf{j}}$
(e) $\mathbf{v} = x^2 y\hat{\mathbf{j}}$ (f) $\mathbf{v} = f(y)\hat{\mathbf{i}} + g(x)\hat{\mathbf{j}}$

3. Are the parentheses needed in (14), i.e., in $\nabla \times (\nabla \times \mathbf{v})$? Explain.

4. $\nabla^2 u = 0$ is called **Laplace's equation**. Show that the following are solutions of Laplace's equation.

(a) $xy + 3z$ (b) $x^2 - y^2 - 2xz$
(c) $x^3 - 3xy^2$ (d) $y^3 - 3x^2 y$
(e) $\ln(x^2 + y^2)$ for $x^2 + y^2 \neq 0$
(f) $\dfrac{1}{x^2 + y^2 + z^2}$ for $x^2 + y^2 + z^2 \neq 0$

5. If $\mathbf{v}$ is any vector, and $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, show that $(\mathbf{v} \cdot \nabla)\mathbf{R} = \mathbf{v}$.

6. Derive the following equations [as we have done in the text for (4)].

(a) equation (5) (b) equation (6)
(c) equation (7) (d) equation (12)
(e) equation (13) (f) equation (14)

7. The following scalar equations occur in fluid mechanics, solid mechanics, and electromagnetics. Re-express them more concisely in terms of vector and vector differential operator notation. Here x, y, z subscripts denote the respective components of a given vector, *not* partial derivatives. To illustrate what we are after, consider the equations

$$\frac{\partial u}{\partial x} = F_x, \qquad \frac{\partial u}{\partial y} = F_y, \qquad \frac{\partial u}{\partial z} = F_z \tag{7.1}$$

If we multiply these equations by $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, respectively, and add them, we obtain

$$\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}} = F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}} + F_z\hat{\mathbf{k}}$$

or, in more concise vector notation, $\nabla u = \mathbf{F}$.

(a) $\dfrac{\partial \sigma}{\partial t} + \dfrac{\partial}{\partial x}(\sigma v_x) + \dfrac{\partial}{\partial y}(\sigma v_y) + \dfrac{\partial}{\partial z}(\sigma v_z) = 0$

(b)
$$\sigma \frac{\partial^2 u_x}{\partial t^2} = \mu\left(\frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2} + \frac{\partial^2 u_x}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial x} + F_x$$
$$\sigma \frac{\partial^2 u_y}{\partial t^2} = \mu\left(\frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2} + \frac{\partial^2 u_y}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial y} + F_y$$
$$\sigma \frac{\partial^2 u_z}{\partial t^2} = \mu\left(\frac{\partial^2 u_z}{\partial x^2} + \frac{\partial^2 u_z}{\partial y^2} + \frac{\partial^2 u_z}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial z} + F_z,$$

where c, μ, λ are constants

(c)
$$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} + \frac{\partial B_x}{\partial t} = 0$$
$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} + \frac{\partial B_y}{\partial t} = 0$$
$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} + \frac{\partial B_z}{\partial t} = 0$$

(d) $$\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} = J_x$$
$$\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} = J_y$$
$$\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} = J_z$$

8. The differential operator $\nabla^4 \equiv \nabla^2\nabla^2$ is known as the **biharmonic** (or **bi-Laplacian**) operator, and the partial differential equation $\nabla^4 u = 0$ is known as the **biharmonic equation**.

(a) If $u = u(x, y)$, show that

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy},$$

where the subscripts denote partial derivatives.
(b) Write out $\nabla^4 u$, in Cartesian coordinates, for the three-dimensional case $u = u(x, y, z)$.

9. (*Caution*) Familiar properties such as $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$, $\mathbf{B} \times \mathbf{A} = -\mathbf{A} \times \mathbf{B}$, and $\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = (\mathbf{A} \times \mathbf{B}) \cdot \mathbf{C}$ are, in general, *not* true if one or more of the vectors is the vector differential operator ∇. Specifically, show that

$$\nabla \cdot \mathbf{v} \neq \mathbf{v} \cdot \nabla, \qquad \nabla \times \mathbf{v} \neq -\mathbf{v} \times \nabla, \qquad \text{and}$$
$$\nabla \cdot (\mathbf{v} \times \mathbf{w}) \neq (\nabla \times \mathbf{v}) \cdot \mathbf{w}.$$

10. Given that

$$\nabla \times \mathbf{v} = -\frac{1}{c}\frac{\partial \mathbf{w}}{\partial t}, \tag{10.1}$$

$$\nabla \times \mathbf{w} = \frac{1}{c}\frac{\partial \mathbf{v}}{\partial t}, \tag{10.2}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{10.3}$$

$$\nabla \cdot \mathbf{w} = 0, \tag{10.4}$$

show that $\mathbf{v}$ and $\mathbf{w}$ satisfy the *wave equations*

$$c^2\nabla^2\mathbf{v} = \frac{\partial^2 \mathbf{v}}{\partial t^2} \qquad \text{and} \qquad c^2\nabla^2\mathbf{w} = \frac{\partial^2 \mathbf{w}}{\partial t^2}.$$

HINT: One of equations (1) to (14) will be useful. Also, note that $\dfrac{\partial}{\partial t}\nabla \cdot \mathbf{u} = \nabla \cdot \dfrac{\partial \mathbf{u}}{\partial t}$ and $\dfrac{\partial}{\partial t}\nabla \times \mathbf{u} = \nabla \times \dfrac{\partial \mathbf{u}}{\partial t}$.

11. Evaluate $\nabla^2\mathbf{v}$ if

(a) $\mathbf{v} = xz^2\hat{\mathbf{i}} + y\sin z\hat{\mathbf{k}}$
(b) $\mathbf{v} = x^2\hat{\mathbf{i}} + y^2\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - 2y\hat{\mathbf{j}} + 3z\hat{\mathbf{k}}$
(d) $\mathbf{v} = \sin(xz)\hat{\mathbf{j}}$

16.7 Non-Cartesian Systems; Div, Grad, Curl, and Laplacian (Optional)

Recall the pattern of the last several sections. Our starting point was the definition

$$\operatorname{div}\mathbf{v} \equiv \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\,dA}{V}\right\}. \tag{1}$$

To obtain a more tractable form, we introduced a reference Cartesian coordinate system and, accordingly, specified $\mathcal{B}$ to be a prism, of dimension Δx by Δy by Δz, the faces of $\mathcal{B}$ being made up of constant-coordinate surfaces: $x =$ constant, $y =$ constant, $z =$ constant. Simplification of the right-hand side of (1) then led to the form

$$\operatorname{div} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}, \tag{2}$$

which we then expressed as

$$\left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}),$$

or $\nabla \cdot \mathbf{v}$, where

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}. \tag{3}$$

With ∇ in hand we then introduced the gradient and curl as $\operatorname{grad} u \equiv \nabla u$ and $\operatorname{curl} \mathbf{v} \equiv \nabla \times \mathbf{v}$.

But Cartesian coordinates are not necessarily the most convenient ones in a given application. For example, in studying the temperature field in a spherical domain, one would certainly be better advised to use spherical coordinates. We begin by considering the cases of plane polar, cylindrical, and spherical coordinates. Or, since the plane polar system is really a special case of the cylindrical system, it will suffice to consider cylindrical and spherical coordinates. In the final subsection, 16.7.3, we give the extension to *any* orthogonal curvilinear coordinate system.

16.7.1. Cylindrical coordinates. Our plan is to derive the ∇ operator, for cylindrical coordinates, from its Cartesian expression (3) by expanding

$$\begin{aligned}\hat{\mathbf{i}} &= \cos\theta \hat{\mathbf{e}}_r - \sin\theta \hat{\mathbf{e}}_\theta,\\ \hat{\mathbf{j}} &= \sin\theta \hat{\mathbf{e}}_r + \cos\theta \hat{\mathbf{e}}_\theta,\\ \hat{\mathbf{k}} &= \hat{\mathbf{e}}_z,\end{aligned} \tag{4}$$

recalling the relations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$, and using chain differentiation. Thus (with $c \equiv \cos\theta$ and $s \equiv \sin\theta$, for brevity),

$$\begin{aligned}\nabla &= \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\\ &= (c\hat{\mathbf{e}}_r - s\hat{\mathbf{e}}_\theta)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial x} + \frac{\partial}{\partial z}\frac{\partial z}{\partial x}\right)\\ &\quad + (s\hat{\mathbf{e}}_r + c\hat{\mathbf{e}}_\theta)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial y} + \frac{\partial}{\partial z}\frac{\partial z}{\partial y}\right)\\ &\quad + (\hat{\mathbf{e}}_z)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial z} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial z} + \frac{\partial}{\partial z}\frac{\partial z}{\partial z}\right).\end{aligned} \tag{5}$$

Recalling from (41) and (42) in Section 13.6 that $r_x = \cos\theta$, $r_y = \sin\theta$, $\theta_x = -\sin\theta/r$, $\theta_y = \cos\theta/r$, noting that $z_x = z_y = 0$ and $z_z = 1$, and using the identity $\cos^2\theta + \sin^2\theta = 1$, (5) reduces to

$$\boxed{\nabla = \hat{\mathbf{e}}_r\frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\mathbf{e}}_z\frac{\partial}{\partial z}.} \tag{6}$$

The ∇ operator is the key. It immediately gives, for the gradient of a scalar field u,

$$\boxed{\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta + \frac{\partial u}{\partial z}\hat{\mathbf{e}}_z,} \tag{7}$$

and with it we can work out the divergence ($\nabla\cdot$), curl ($\nabla\times$), and Laplacian (∇^2) of a given field. In doing so, it will be crucial to remember that the cylindrical coordinate base vectors are not constants. Specifically, recall from Section 14.6 that $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ are functions of θ and $\hat{\mathbf{e}}_z$ is a constant, with the only nonzero derivatives being

$$\frac{d\hat{\mathbf{e}}_r}{d\theta}=\hat{\mathbf{e}}_\theta \qquad\text{and}\qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta}=-\hat{\mathbf{e}}_r. \tag{8}$$

For example, given the vector field $\mathbf{v}=v_r\hat{\mathbf{e}}_r+v_\theta\hat{\mathbf{e}}_\theta+v_z\hat{\mathbf{e}}_z$, consider its divergence

$$\nabla\cdot\mathbf{v}=\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}+\hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial\theta}+\hat{\mathbf{e}}_z\frac{\partial}{\partial z}\right)\cdot\left(v_r\hat{\mathbf{e}}_r+v_\theta\hat{\mathbf{e}}_\theta+v_z\hat{\mathbf{e}}_z\right). \tag{9}$$

Carrying out the dot product on the right-hand side yields nine terms: the first term in ∇ dotted with the first term in $\mathbf{v}$, the first with the second, the first with the third, the second with the first, and so on. As representative, consider the first term dotted with the first,

$$\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right). \tag{10}$$

Notice carefully that the latter is ambiguous because there are two operations to carry out, a dot product and a derivative, and we are not told which to do first and which to do second. In the case of (10) we get the same result either way, but that is not true for all of the terms in (9). We state, without proof, that the correct answer in such cases, as could be verified by working the problem in a different way, is always obtained if we *do the differentation first and then the dot product (or cross product if we are computing the curl rather than the divergence)*. Thus, to work out (10) we write

$$\begin{aligned}\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right)&=\hat{\mathbf{e}}_r\cdot\frac{\partial}{\partial r}\left(v_r\hat{\mathbf{e}}_r\right)=\hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_r}{\partial r}\hat{\mathbf{e}}_r+v_r\frac{\partial\hat{\mathbf{e}}_r}{\partial r}\right)\\&=\frac{\partial v_r}{\partial r}\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_r+v_r\hat{\mathbf{e}}_r\cdot\mathbf{0}=\frac{\partial v_r}{\partial r}\end{aligned} \tag{11}$$

since $\partial\hat{\mathbf{e}}_r/\partial r=\mathbf{0}$. Observe that in the first equality, in (11), we have slid the $\partial/\partial r$ past the dot. That step gives $\hat{\mathbf{e}}_r\cdot\partial(v_r\hat{\mathbf{e}}_r)/\partial r$, which is unambiguous; it instructs us to *first* work out the derivative $\partial(v_r\hat{\mathbf{e}}_r)/\partial r$ and *then* the dot product, as explained above in italicized type.

Really, the only terms requiring such special care are the second term dotted with the first (because $\partial\hat{\mathbf{e}}_r/\partial\theta\neq\mathbf{0}$) and the second term dotted with the second (because $\partial\hat{\mathbf{e}}_\theta/\partial\theta\neq\mathbf{0}$):

$$\begin{aligned}\left(\hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial\theta}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right)&=\hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\frac{\partial}{\partial\theta}\left(v_r\hat{\mathbf{e}}_r\right)\\&=\hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_r}{\partial\theta}\hat{\mathbf{e}}_r+v_r\frac{\partial\hat{\mathbf{e}}_r}{\partial\theta}\right)\\&=\hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_r}{\partial\theta}\hat{\mathbf{e}}_r+v_r\,\hat{\mathbf{e}}_\theta\right)=\frac{v_r}{r}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\left(\hat{\mathbf{e}}_\theta \frac{1}{r}\frac{\partial}{\partial\theta}\right)\cdot(v_\theta\hat{\mathbf{e}}_\theta) &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\frac{\partial}{\partial\theta}(v_\theta\hat{\mathbf{e}}_\theta)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_\theta}{\partial\theta}\hat{\mathbf{e}}_\theta + v_\theta\frac{\partial\hat{\mathbf{e}}_\theta}{\partial\theta}\right)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_\theta}{\partial\theta}\hat{\mathbf{e}}_\theta - v_\theta\,\hat{\mathbf{e}}_r\right) = \frac{1}{r}\frac{\partial v_\theta}{\partial\theta}. \qquad (13)\end{aligned}$$

Of the nine terms that result from (9) we have worked out three, in (11)–(13). Let us work out just one more, say the first term in $\boldsymbol{\nabla}$ dotted with the third in $\mathbf{v}$, and then state the final result:

$$\begin{aligned}\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot(v_z\hat{\mathbf{e}}_z) &= \hat{\mathbf{e}}_r\cdot\frac{\partial}{\partial r}(v_z\hat{\mathbf{e}}_z) = \hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_z}{\partial r}\hat{\mathbf{e}}_z + v_z\frac{\partial\hat{\mathbf{e}}_z}{\partial z}\right)\\ &= \hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_z}{\partial r}\hat{\mathbf{e}}_z + \mathbf{0}\right) = 0 + 0 = 0. \qquad (14)\end{aligned}$$

In fact, five of the nine terms give 0, and the final result is

$$\boldsymbol{\nabla}\cdot\mathbf{v} = \frac{\partial v_r}{\partial r} + \frac{1}{r}v_r + \frac{1}{r}\frac{\partial v_\theta}{\partial\theta} + \frac{\partial v_z}{\partial z}, \qquad (15)$$

which is often expressed, more compactly, as

$$\boxed{\boldsymbol{\nabla}\cdot\mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial}{\partial\theta}v_\theta + \frac{\partial}{\partial z}v_z.} \qquad (16)$$

Proceeding in similar fashion, one obtains

$$\boxed{\boldsymbol{\nabla}\times\mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial\theta} - \frac{\partial v_\theta}{\partial z}\right)\hat{\mathbf{e}}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\hat{\mathbf{e}}_\theta + \frac{1}{r}\left(\frac{\partial(rv_\theta)}{\partial r} - \frac{\partial v_r}{\partial\theta}\right)\hat{\mathbf{e}}_z} \qquad (17)$$

and

$$\boxed{\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2} + \frac{\partial^2 u}{\partial z^2}.} \qquad (18)$$

Thus, the gradient and the Laplacian of any scalar field $u(r,\theta,z)$ are given by (7) and (18), respectively, and the divergence and curl of any vector field $\mathbf{v}(r,\theta,z)$ are given by (16) and (17), respectively.

EXAMPLE 1. Consider the scalar field $u = rz^2\sin\theta$, and the vector field $\mathbf{v} = z^2\hat{\mathbf{e}}_r - 2\hat{\mathbf{e}}_\theta + r\hat{\mathbf{e}}_z$. Then (7) and (18) give

$$\boldsymbol{\nabla}u = z^2\sin\theta\hat{\mathbf{e}}_r + z^2\cos\theta\hat{\mathbf{e}}_\theta + 2rz\sin\theta\hat{\mathbf{e}}_z \qquad (19)$$

and

$$\nabla^2 u = 0 + \frac{z^2}{r}\sin\theta - \frac{rz^2}{r^2}\sin\theta + 2r\sin\theta = 2r\sin\theta, \tag{20}$$

respectively. To use (16) and (17) to work out $\nabla \cdot \mathbf{v}$ and $\nabla \times \mathbf{v}$ we need, first, to identify the components $v_r(r, \theta, z)$, $v_\theta(r, \theta, z)$, and $v_z(r, \theta, z)$ of $\mathbf{v}$. Since $\mathbf{v} = v_r\hat{\mathbf{e}}_r + v_\theta\hat{\mathbf{e}}_\theta + v_z\hat{\mathbf{e}}_z = z^2\hat{\mathbf{e}}_r - 2\hat{\mathbf{e}}_\theta + r\hat{\mathbf{e}}_z$, we see that $v_r = z^2$, $v_\theta = -2$, and $v_z = r$. Thus,

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rz^2) + \frac{1}{r}\frac{\partial}{\partial \theta}(-2) + \frac{\partial}{\partial z}(r) = \frac{z^2}{r} \tag{21}$$

from (16), and

$$\begin{aligned}\nabla \times \mathbf{v} &= (0-0)\hat{\mathbf{e}}_r + (2z-1)\hat{\mathbf{e}}_\theta + \frac{1}{r}(-2-0)\hat{\mathbf{e}}_z \\ &= (2z-1)\hat{\mathbf{e}}_\theta - \frac{2}{r}\hat{\mathbf{e}}_z\end{aligned} \tag{22}$$

from (17).

COMMENT. In this case $u = rz^2\sin\theta$ is readily re-expressed in terms of x, y, z since $r\sin\theta$ is y. Thus, $u = yz^2$ so

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0 + 0 + 2y = 2y,$$

which does agree with (20) because $y = r\sin\theta$. ▮

As noted above, we need not derive these operators for the case of plane polar coordinates because they merely follow from (7), (16), (17), and (18). That is, if

$$u = u(r, \theta) \qquad \text{and} \qquad \mathbf{v} = v_r(r,\theta)\hat{\mathbf{e}}_r + v_\theta(r,\theta)\hat{\mathbf{e}}_\theta,$$

then all z derivatives are zero and $v_z = 0$ too, so those formulas merely reduce to

$$\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta, \tag{23}$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}, \tag{24}$$

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial}{\partial \theta}v_\theta, \tag{25}$$

$$\nabla \times \mathbf{v} = \frac{1}{r}\left(\frac{\partial(rv_\theta)}{\partial r} - \frac{\partial v_r}{\partial \theta}\right)\hat{\mathbf{e}}_z. \tag{26}$$

16.7.2. Spherical coordinates. Here we consider scalar fields u and vector fields

$$\mathbf{v} = v_\rho\hat{\mathbf{e}}_\rho + v_\phi\hat{\mathbf{e}}_\phi + v_\theta\hat{\mathbf{e}}_\theta, \tag{27}$$

where u, v_ρ, v_ϕ, and v_θ are functions of the spherical coordinates ρ, ϕ, θ that were introduced in Section14.6, the latter being related to x, y, z according to

$$\begin{aligned} x &= \rho \sin\phi \cos\theta, \\ y &= \rho \sin\phi \sin\theta, \\ z &= \rho \cos\phi. \end{aligned} \tag{28}$$

Derivation of ∇, $\nabla \cdot \mathbf{v}$, $\nabla^2 u$, and so on, follows exactly the same lines as indicated above for cylindrical coordinates. For brevity we merely state the main results and leave derivations for the exercises.

$$\boxed{\nabla = \hat{\mathbf{e}}_\rho \frac{\partial}{\partial \rho} + \hat{\mathbf{e}}_\phi \frac{1}{\rho}\frac{\partial}{\partial \phi} + \hat{\mathbf{e}}_\theta \frac{1}{\rho \sin\phi}\frac{\partial}{\partial \theta},} \tag{29}$$

$$\boxed{\nabla u = \frac{\partial u}{\partial \rho}\hat{\mathbf{e}}_\rho + \frac{1}{\rho}\frac{\partial u}{\partial \phi}\hat{\mathbf{e}}_\phi + \frac{1}{\rho \sin\phi}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta,} \tag{30}$$

$$\boxed{\nabla^2 u = \frac{1}{\rho^2}\left[\frac{\partial}{\partial \rho}\left(\rho^2 \frac{\partial u}{\partial \rho}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi \frac{\partial u}{\partial \phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}\right],} \tag{31}$$

$$\boxed{\nabla \cdot \mathbf{v} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}(\rho^2 v_\rho) + \frac{1}{\rho \sin\phi}\frac{\partial}{\partial \phi}(v_\phi \sin\phi) + \frac{1}{\rho \sin\phi}\frac{\partial v_\theta}{\partial \theta},} \tag{32}$$

$$\boxed{\begin{aligned} \nabla \times \mathbf{v} = {} & \frac{1}{\rho \sin\phi}\left(\frac{\partial}{\partial \phi}(v_\theta \sin\phi) - \frac{\partial v_\phi}{\partial \theta}\right)\hat{\mathbf{e}}_\rho \\ & + \frac{1}{\rho}\left(\frac{1}{\sin\phi}\frac{\partial v_\rho}{\partial \theta} - \frac{\partial(\rho v_\theta)}{\partial \rho}\right)\hat{\mathbf{e}}_\phi + \frac{1}{\rho}\left(\frac{\partial(\rho v_\phi)}{\partial \rho} - \frac{\partial v_\rho}{\partial \phi}\right)\hat{\mathbf{e}}_\theta. \end{aligned}} \tag{33}$$

EXAMPLE 2. Given $u = \rho^2 \sin^2\phi \sin^2\theta$ and $\mathbf{v} = \rho \hat{\mathbf{e}}_\rho$, use the given formulas to determine $\nabla^2 u$ and $\nabla \cdot \mathbf{v}$. From (31),

$$\begin{aligned} \nabla^2 u = {} & \frac{1}{\rho^2}\left[\frac{\partial}{\partial \rho}\left(2\rho^3 \sin^2\phi \sin^2\theta\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(2\rho^2 \sin^2\phi \cos\phi \sin^2\theta\right)\right. \\ & \left. + \frac{1}{\sin^2\phi} 2\rho^2 \sin^2\phi(\cos^2\theta - \sin^2\theta)\right] = 2, \end{aligned}$$

where we omit the final steps, for brevity, since they are straightforward. And from (32), with $v_\rho = \rho$, $v_\phi = 0$, and $v_\theta = 0$,

$$\nabla \cdot \mathbf{v} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\rho^3 = 3.$$

COMMENT. We see from (28) that $u = \rho^2 \sin^2 \phi \sin^2 \theta$ is readily expressed in terms of Cartesian coordinates as $u = y^2$, and $\mathbf{v} = \rho \hat{\mathbf{e}}_\rho$ is simply the position vector $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$. Thus,

$$\nabla^2 u = \frac{\partial^2}{\partial x^2}(y^2) + \frac{\partial^2}{\partial y^2}(y^2) + \frac{\partial^2}{\partial z^2}(y^2) = 0 + 2 + 0 = 2$$

and

$$\nabla \cdot \mathbf{v} = \frac{\partial}{\partial x}(x) + \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial z}(z) = 1 + 1 + 1 = 3,$$

both of which results agree with the results found above using (31) and (32). ▮

A word of CAUTION regarding the use of determinants for the representation of $\nabla \times \mathbf{v}$ in non-Cartesian coordinate systems: Recall that it turns out that the cross product $\mathbf{u} \times \mathbf{v}$ of two vectors $\mathbf{u} = u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}$ and $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$ can be expressed as the determinant

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{vmatrix}, \tag{34}$$

as can be verified by working out $(u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}) \times (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}})$ term by term, expanding the determinant in (34), and verifying that the two results are identical.

Similarly for *any* right-handed set of orthonormal base vectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$, where by right-handed we mean that $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3$, $\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3 = \hat{\mathbf{e}}_1$, and $\hat{\mathbf{e}}_3 \times \hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_2$. That is, if $\mathbf{u} = u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3$ and $\mathbf{v} = v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3$, then

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}, \tag{35}$$

as can be verified by working out $(u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3) \times (v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3)$ term by term, expanding the determinant in (35), and verifying that the two results are identical. For instance, $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ might be the right-handed cylindrical coordinate base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$, or the right-handed spherical coordinate base vectors $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$.

However, if the $\mathbf{u}$ vector is not an ordinary vector but is the vector differential operator ∇, then the determinant form for $\nabla \times \mathbf{v}$ gives INCORRECT results for non-Cartesian systems. For example, the cylindrical coordinate formula

$$\nabla \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{e}}_r & \hat{\mathbf{e}}_\theta & \hat{\mathbf{e}}_z \\ \partial/\partial r & (1/r)\partial/\partial\theta & \partial/\partial z \\ v_r & v_\theta & v_z \end{vmatrix}$$

is INCORRECT, as can be seen by expanding the determinant and comparing the result with the correct result given by (17). A generalized determinant formula for $\nabla \times \mathbf{v}$ that *is* correct will be given below by (43).

16.7.3. General orthogonal coordinates. Let us give the extension of these results to any right-handed orthogonal curvilinear coordinate system with coordinates q_1, q_2, q_3 and corresponding base vectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$. Let q_1, q_2, q_3 be related to the reference Cartesian coordinates according to

$$\begin{aligned} x &= x(q_1, q_2, q_3), \\ y &= y(q_1, q_2, q_3), \\ z &= z(q_1, q_2, q_3). \end{aligned} \tag{36}$$

For instance, q_1, q_2, q_3 might be the cylindrical coordinates r, θ, z, respectively.

The key is to identify the quantities h_1, h_2, h_3 in

$$d\mathbf{R} = h_1 dq_1\, \hat{\mathbf{e}}_1 + h_2 dq_2\, \hat{\mathbf{e}}_2 + h_3 dq_3\, \hat{\mathbf{e}}_3, \tag{37}$$

which are given by

$$\boxed{\begin{aligned} h_1 &= \sqrt{\left(\frac{\partial x}{\partial q_1}\right)^2 + \left(\frac{\partial y}{\partial q_1}\right)^2 + \left(\frac{\partial z}{\partial q_1}\right)^2}, \\ h_2 &= \sqrt{\left(\frac{\partial x}{\partial q_2}\right)^2 + \left(\frac{\partial y}{\partial q_2}\right)^2 + \left(\frac{\partial z}{\partial q_2}\right)^2}, \\ h_3 &= \sqrt{\left(\frac{\partial x}{\partial q_3}\right)^2 + \left(\frac{\partial y}{\partial q_3}\right)^2 + \left(\frac{\partial z}{\partial q_3}\right)^2}. \end{aligned}} \tag{38}$$

With the h_j's determined from (38), we have

$$\boxed{\nabla = \hat{\mathbf{e}}_1 \frac{1}{h_1}\frac{\partial}{\partial q_1} + \hat{\mathbf{e}}_2 \frac{1}{h_2}\frac{\partial}{\partial q_2} + \hat{\mathbf{e}}_3 \frac{1}{h_3}\frac{\partial}{\partial q_3},} \tag{39}$$

$$\boxed{\nabla u = \frac{1}{h_1}\frac{\partial u}{\partial q_1}\hat{\mathbf{e}}_1 + \frac{1}{h_2}\frac{\partial u}{\partial q_2}\hat{\mathbf{e}}_2 + \frac{1}{h_3}\frac{\partial u}{\partial q_3}\hat{\mathbf{e}}_3,} \tag{40}$$

$$\boxed{\nabla^2 u = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}\left(\frac{h_2 h_3}{h_1}\frac{\partial u}{\partial q_1}\right) + \frac{\partial}{\partial q_2}\left(\frac{h_1 h_3}{h_2}\frac{\partial u}{\partial q_2}\right) + \frac{\partial}{\partial q_3}\left(\frac{h_1 h_2}{h_3}\frac{\partial u}{\partial q_3}\right)\right],} \tag{41}$$

$$\boxed{\nabla \cdot \mathrm{v} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}(h_2 h_3 v_1) + \frac{\partial}{\partial q_2}(h_1 h_3 v_2) + \frac{\partial}{\partial q_3}(h_1 h_2 v_3)\right],} \tag{42}$$

and

$$\boxed{\nabla \times \mathrm{v} = \frac{1}{h_1 h_2 h_3}\begin{vmatrix} h_1\hat{\mathrm{e}}_1 & h_2\hat{\mathrm{e}}_2 & h_3\hat{\mathrm{e}}_3 \\ \dfrac{\partial}{\partial q_1} & \dfrac{\partial}{\partial q_2} & \dfrac{\partial}{\partial q_3} \\ h_1 v_1 & h_2 v_2 & h_3 v_3 \end{vmatrix},} \tag{43}$$

where v_1, v_2, v_3 denote the $\hat{\mathrm{e}}_1$, $\hat{\mathrm{e}}_2$, $\hat{\mathrm{e}}_3$ components of v, respectively; that is, $\mathrm{v} = v_1\hat{\mathrm{e}}_1 + v_2\hat{\mathrm{e}}_2 + v_3\hat{\mathrm{e}}_3$.

EXAMPLE 3. In cylindrical coordinates, let us identify q_1, q_2, q_3 as r, θ, z, respectively, although the order will not matter provided that q_1, q_2, q_3 is a right-handed system. The relations (36) are

$$\begin{aligned} x &= r\cos\theta = q_1 \cos q_2, \\ y &= r\sin\theta = q_1 \sin q_2, \\ z &= z \qquad\;\; = q_3 \end{aligned}$$

and putting these into (38) gives $h_1 = 1$, $h_2 = q_1$, $h_3 = 1$. With these h_j's, (39)–(43) do give results that are the same as (6), (7), (18), (16), and (17), respectively (Exercise 10). ∎

Closure. The key, in obtaining the divergence, gradient, curl, and Laplacian operators for a non-Cartesian coordinate system is to obtain the ∇ operator. That can be done from its Cartesian form, (3), by expressing $x, y, z, \hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ in terms of the desired coordinates and base vectors. With ∇ in hand, the gradient of a scalar field u is given immediately as ∇u. Knowing the derivatives of the non-Cartesian base vectors, we can also work out expressions for $\nabla \cdot \mathrm{v}$, $\nabla \times \mathrm{v}$, and $\nabla^2 u$. The results, for cylindrical and spherical coordinates, are given in this section and framed for emphasis. The key point to keep in mind, in deriving formulas for $\nabla \cdot \mathrm{v}$ and $\nabla \times \mathrm{v}$, is that the derivatives within the ∇ act not only upon the scalar components within v but also upon the *base vectors* within v, and the derivatives of the base vectors are not all zero for a non-Cartesian system, in general. Similarly, the derivatives within the first ∇, in $\nabla^2 u = \nabla \cdot \nabla u$, act not only upon u but also upon the base vectors within the second ∇. Of course it should be appreciated that once the results (6), (7), (16)–(18), and (30)–(33) are in hand we can simply use those formulas without dealing with the above-mentioned details.

EXERCISES 16.7

Cylindrical coordinates:

1. Evaluate ∇u, $\nabla^2 u$, $\nabla \cdot \mathbf{v}$, and $\nabla \times \mathbf{v}$, using the relevant formulas (7), (16), (17), and (18).

(a) $u = r, \quad \mathbf{v} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$
(b) $u = r\sin 2\theta, \quad \mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(c) $u = r^2\sin\theta, \quad \mathbf{v} = z^2\hat{\mathbf{e}}_r$
(d) $u = z^3 r, \quad \mathbf{v} = \hat{\mathbf{e}}_\theta$
(e) $u = 1/r, \quad \mathbf{v} = r\sin\theta\hat{\mathbf{e}}_z$
(f) $u = 6, \quad \mathbf{v} = z^2\hat{\mathbf{e}}_r + (r^2 z + \cos^2\theta)\hat{\mathbf{e}}_z$
(g) $u = r^3\cos\theta, \quad \mathbf{v} = 6\hat{\mathbf{e}}_r - 3\hat{\mathbf{e}}_\theta - 2\hat{\mathbf{e}}_z$
(h) $u = r^2 + z^2, \quad \mathbf{v} = rz\hat{\mathbf{e}}_r + r^2 z^2\cos\theta\hat{\mathbf{e}}_\theta$

2. Evaluate $\nabla \cdot \mathbf{v}$ two different ways, and show that the results are in agreement with each other: first, use (16); second, re-express $\mathbf{v}$ in terms of $x, y, z, \hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ and use the formula $\nabla \cdot \mathbf{v} = \partial v_x/\partial x + \partial v_y/\partial y + \partial v_z/\partial z$.

(a) $\mathbf{v} = \hat{\mathbf{e}}_r$ (b) $\mathbf{v} = \hat{\mathbf{e}}_\theta$
(c) $\mathbf{v} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$ (d) $\mathbf{v} = \cos^2\theta\hat{\mathbf{e}}_\theta$
(e) $\mathbf{v} = r^2 z\sin\theta\hat{\mathbf{e}}_z$ (f) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(g) $\mathbf{v} = r\hat{\mathbf{e}}_\theta$ (h) $\mathbf{v} = r^3\hat{\mathbf{e}}_r$
(i) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_z$

3. Carry out steps analogous to those followed in (9)–(16), in deriving the formula (16) for $\nabla \cdot \mathbf{v}$, to derive

(a) the formula (26) for $\nabla \times \mathbf{v}$ in plane polar coordinates
(b) the formula (17) for $\nabla \times \mathbf{v}$ in cylindrical coordinates

4. Derive the formula (18) for $\nabla^2 u$

(a) by writing $\nabla^2 u = \nabla \cdot \nabla u$ and using (7) and (16)
(b) by beginning with the Cartesian version

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2},$$

introducing the relations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$, and using chain differentiation. HINT: $\partial^2/\partial x^2$, for example, is a double application of the operator

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial x} + \frac{\partial}{\partial z}\frac{\partial z}{\partial x}$$
$$= \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}.$$

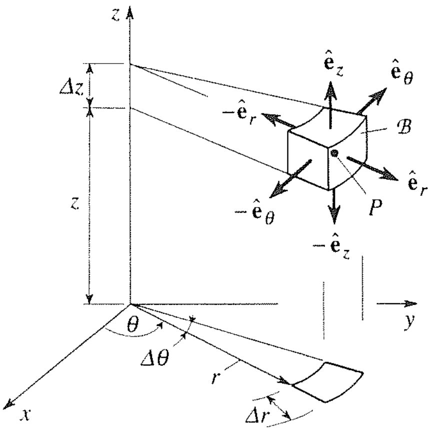
Similarly,

$$\frac{\partial}{\partial y} = \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial z} = \frac{\partial}{\partial z}.$$

5. (a) Derive the formula (16) for $\nabla \cdot \mathbf{v}$ using the limit definition

$$\operatorname{div}\mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}, \tag{5.1}$$

and choosing $\mathcal{B}$ as shown in the accompanying figure, bounded by constant-r, constant-θ, constant-z surfaces. HINT: The analogous derivation, for Cartesian coordinates, is given in Section 16.3.



(b) Derive the formula (9) for ∇u using the limit definition

$$\operatorname{grad} u = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\,u\,dA}{V}\right\}, \tag{5.2}$$

and choosing $\mathcal{B}$ as shown in the figure.
(c) Derive the formula (17) for $\nabla\times\mathbf{v}$ using the limit definition

$$\operatorname{curl}\mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\times\mathbf{v}\,dA}{V}\right\}, \tag{5.3}$$

and choosing $\mathcal{B}$ as shown in the figure.

6. Show that the given u is a solution of the *Laplace equation* $\nabla^2 u = 0$.

(a) $u = r^n\sin n\theta$ for any integer n (for $r \neq 0$ if n is negative)
(b) $u = r^n\cos n\theta$ for any integer n (for $r \neq 0$ if n is negative)
(c) $u = \ln r \quad (r \neq 0)$
(d) $u = \dfrac{1}{\sqrt{r^2 + z^2}} \quad (r^2 + z^2 \neq 0)$

7. Besides $\nabla^2 u$, we can also have $\nabla^2\mathbf{v}$, the Laplacian operating on a vector field. Rather than ask you to work out $\nabla^2\mathbf{v}$ in cylindrical coordinates, because it is a bit messy, we ask you to derive the plane polar coordinate result

$$\nabla^2 \mathbf{v} = \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_r}{\partial r}\right) + \frac{1}{r^2}\left(\frac{\partial^2 v_r}{\partial \theta^2} - 2\frac{\partial v_\theta}{\partial \theta} - v_r\right)\right]\hat{\mathbf{e}}_r + \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_\theta}{\partial r}\right) + \frac{1}{r^2}\left(\frac{\partial^2 v_\theta}{\partial \theta^2} + 2\frac{\partial v_r}{\partial \theta} - v_\theta\right)\right]\hat{\mathbf{e}}_\theta \tag{7.1}$$

by working out

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right)(v_r(r,\theta)\hat{\mathbf{e}}_r + v_\theta(r,\theta)\hat{\mathbf{e}}_\theta)$$

term by term and remembering that $\partial\hat{\mathbf{e}}_r/\partial r = \partial\hat{\mathbf{e}}_\theta/\partial r = \mathbf{0}$, $\partial\hat{\mathbf{e}}_r/\partial\theta = \hat{\mathbf{e}}_\theta$, and $\partial\hat{\mathbf{e}}_\theta/\partial\theta = -\hat{\mathbf{e}}_r$.

8. Use (7.1), above, to evaluate $\nabla^2\mathbf{v}$.

(a) $\mathbf{v} = r^2\hat{\mathbf{e}}_r$ (b) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(c) $\mathbf{v} = \hat{\mathbf{e}}_\theta$ (d) $\mathbf{v} = r\cos\theta\hat{\mathbf{e}}_r - r^2\sin 3\theta\hat{\mathbf{e}}_\theta$

9. Write out the *biharmonic equation* $\nabla^4 u = 0$, where $\nabla^4 \equiv \nabla^2\nabla^2$, in both Cartesian coordinates and cylindrical coordinates.

10. Verify that for the cylindrical coordinates $q_1 = r$, $q_2 = \theta$, $q_3 = z$, equations (39)–(43) do give the same results as (6), (7), (18), (16), and (17) as claimed in Example 3.

Spherical coordinates:

11. Evaluate ∇u, $\nabla^2 u$, $\nabla\cdot\mathbf{v}$, and $\nabla\times\mathbf{v}$, using the relevant formulas (30)–(33).

(a) $u = \rho$, $\mathbf{v} = \rho\hat{\mathbf{e}}_\rho$
(b) $u = \rho^2$, $\mathbf{v} = 3\hat{\mathbf{e}}_\phi$
(c) $u = \sin\phi$, $\mathbf{v} = \rho\hat{\mathbf{e}}_\phi$
(d) $u = \rho^2\sin\theta$, $\mathbf{v} = \hat{\mathbf{e}}_\rho + \hat{\mathbf{e}}_\phi + \hat{\mathbf{e}}_\theta$
(e) $u = \rho\cos\theta$, $\mathbf{v} = \rho^3\hat{\mathbf{e}}_\theta$
(f) $u = \cos\phi$, $\mathbf{v} = 6\hat{\mathbf{e}}_\theta$

12. Using the expression

$$\nabla = \hat{\mathbf{e}}_\rho\frac{\partial}{\partial\rho} + \hat{\mathbf{e}}_\phi\frac{1}{\rho}\frac{\partial}{\partial\phi} + \hat{\mathbf{e}}_\theta\frac{1}{\rho\sin\phi}\frac{\partial}{\partial\theta},$$

and, as needed, the expressions for the space derivatives of $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ given in Section 14.6.3, derive

(a) the formula (32) for $\nabla\cdot\mathbf{v}$
(b) the formula (33) for $\nabla\times\mathbf{v}$
(c) the formula (31) for $\nabla^2 u$

13. Find values of α, if any, such that $u = \rho^\alpha$ is a solution of the *Laplace equation* $\nabla^2 u = 0$.

14. For spherical polars, x, y, z are related to ρ, ϕ, θ according to (28). Letting q_1, q_2, q_3 be ρ, ϕ, θ, respectively, show that (39)–(43) give expressions for ∇, ∇u, $\nabla^2 u$, $\nabla\cdot\mathbf{v}$, and $\nabla\times\mathbf{v}$ that agree with (29)–(33).

15. The vector equation

$$\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\sigma}\nabla p, \tag{15.1}$$

where $\mathbf{v}$ is the vector velocity field, p is the scalar pressure field, t is the time, and σ is the fluid mass density, is known in fluid mechanics as the *equation of motion* (for a fluid having no viscosity) since it is the fluid mechanics version of Newton's second law of motion $\mathbf{F} = m\mathbf{a}$ for a particle of constant mass m. Write the three scalar equations that follow from (15.1), in

(a) Cartesian coordinates x, y, z
(b) cylindrical coordinates r, θ, z
(c) spherical coordinates ρ, ϕ, θ

16.8 Divergence Theorem

16.8.1. Divergence theorem. In Sections 16.1–16.7 we introduce the quantities that we will need – div, grad, curl, Laplacian, and so on. In this section we begin to use those quantities to obtain important results. In particular, we derive the Gauss divergence theorem and use it to derive well known equations in fluid mechanics and heat transfer.*

*Generally attributed to *Gauss*, the divergence theorem was obtained independently by the Rus-

THEOREM 16.8.1 *Divergence Theorem*
Let $\mathcal{V}$ be a closed region in 3-space, the boundary of which is a piecewise smooth orientable surface $\mathcal{S}$, and let $\mathbf{v}$ be a vector field that is defined and C^1 in $\mathcal{V}$. If $\hat{\mathbf{n}}$ denotes the outward unit normal on $\mathcal{S}$, then

$$\boxed{\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA.} \tag{1}$$

Outline of Proof: We begin by proving (1) for the simple case where $\mathcal{V}$ is a rectangular prism $x_1 \le x \le x_2$, $y_1 \le y \le y_2$, $z_1 \le z \le z_2$. Write out

$$\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV = \int_{\mathcal{V}} \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) dV. \tag{2}$$

Consider the first term on the right:

$$\begin{aligned}
\int_{\mathcal{V}} \frac{\partial v_x}{\partial x}\, dV &= \int_{z_1}^{z_2} \int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} \frac{\partial v_x}{\partial x}\, dx \right) dy\, dz \\
&= \int_{z_1}^{z_2} \int_{y_1}^{y_2} v_x(x_2, y, z)\, dy\, dz - \int_{z_1}^{z_2} \int_{y_1}^{y_2} v_x(x_1, y, z)\, dy\, dz \\
&= \int \int_{x = x_2 \text{ face}} \mathbf{v} \cdot \hat{\mathbf{i}}\, dy\, dz + \int \int_{x = x_1 \text{ face}} \mathbf{v} \cdot (-\hat{\mathbf{i}})\, dy\, dz.
\end{aligned} \tag{3}$$

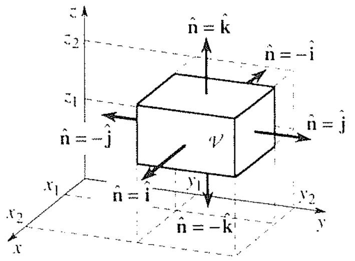


Figure 1. The prism $\mathcal{V}$.

Since $\hat{\mathbf{n}} = +\hat{\mathbf{i}}$ on the $x = x_2$ face and $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ on the $x = x_1$ face (Fig. 1), (3) amounts to

$$\int_{\mathcal{V}} \frac{\partial v_x}{\partial x}\, dV = \int_{x_1, x_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA. \tag{4}$$

Similarly,

$$\begin{aligned}
\int_{\mathcal{V}} \frac{\partial v_y}{\partial y}\, dV &= \int_{z_1}^{z_2} \int_{x_1}^{x_2} \left(\int_{y_1}^{y_2} \frac{\partial v_y}{\partial y}\, dy \right) dx\, dz \\
&= \text{etc.} = \int_{y_1, y_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA
\end{aligned} \tag{5}$$

and

$$\int_{\mathcal{V}} \frac{\partial v_z}{\partial z}\, dV = \int_{z_1, z_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA. \tag{6}$$

sian mathematician *Michel Ostrogradsky* (1801–1861), and was published in the memoirs of the St. Petersburg Academy of Sciences in 1831.

Then, adding (4)–(6) gives (1). We need to justify using different orders of integration in the three volume integrals; that is, to integrate $\partial v_x/\partial x$, $\partial v_y/\partial y$, $\partial v_z/\partial z$ we integrated first with respect to x, y, z, respectively. Since the theorem assumes $\mathbf{v}$ to be C^1, it follows that the integrand [written out in the right-hand side of (2)] is continuous, so the integral has a unique value, independent of the order of integration. [See equations (8) and (24), and the associated discussion, in Section 15.3.]

The preceding proof can be generalized to any closed region $\mathcal{V}$ in 3-space, the boundary of which is a piecewise smooth orientable surface. A partial generalization is covered in the exercises. Let us use this space instead to outline a different approach which has the advantage of flowing directly from the limit definition of the divergence,

$$\operatorname{div}\mathbf{v}(P) = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_{\mathcal{S}}\hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}, \tag{7}$$

but understand that our presentation will be heuristic, not rigorous. To begin, let us break $\mathcal{V}$ into N subregions, or cells, where N is very large; one might think of $\mathcal{V}$ as being filled with millions of fish eggs. Let us number them, from 1 to N, call the jth cell $\mathcal{B}_j$, with surface $\mathcal{S}_j$, volume dV_j, and with some point in $\mathcal{B}_j$ designated as P_j. Some such cells are sketched in Fig. 2a, at an exaggerated scale so we can label them. Let us write down (7) for each cell. Dropping the limit, but reminding ourselves of it by writing dV in place of V, we have the N equations

$$\begin{aligned}\operatorname{div}\mathbf{v}(P_1)dV_1 &= \int_{\mathcal{S}_1}\hat{\mathbf{n}}_1\cdot\mathbf{v}\,dA\\ &\vdots\\ \operatorname{div}\mathbf{v}(P_N)dV_N &= \int_{\mathcal{S}_N}\hat{\mathbf{n}}_N\cdot\mathbf{v}\,dA.\end{aligned} \tag{8}$$

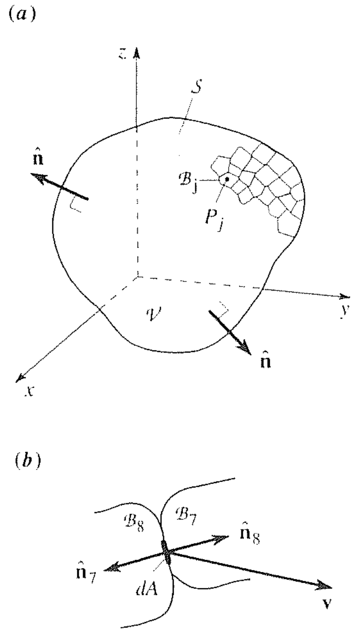


Figure 2. Partition of $\mathcal{V}$.

Adding these, the sum of the left-hand sides is $\int_{\mathcal{V}}\operatorname{div}\mathbf{v}\,dV$, namely, the left-hand member of (1). Interpreting the sum of the right-hand sides of (8) is a bit trickier. Consider two adjacent cells, say $\mathcal{B}_7$ and $\mathcal{B}_8$ (Fig. 2b), and observe that the contributions $\hat{\mathbf{n}}_7\cdot\mathbf{v}\,dA$ to $\int_{\mathcal{S}_7}\hat{\mathbf{n}}_7\cdot\mathbf{v}\,dA$ and $\hat{\mathbf{n}}_8\cdot\mathbf{v}\,dA$ to $\int_{\mathcal{S}_8}\hat{\mathbf{n}}_8\cdot\mathbf{v}\,dA$ exactly cancel because $\hat{\mathbf{n}}_8 = -\hat{\mathbf{n}}_7$. Such cancellation occurs along all of the internal boundaries. Only the contributions from the dA elements lying on the outer surface $\mathcal{S}$ survive because these have no neighbors and hence suffer no such cancellation. Thus, the sum of the right-hand sides of (8) is $\int_{\mathcal{S}}\hat{\mathbf{n}}\cdot\mathbf{v}\,dA$, namely, the right-hand member of (1). ■

From the preceding heuristic proof we can see that the divergence theorem (1) [more precisely, the divergence theorem *formula* (1) since (1) is a formula, not a theorem] is actually a macro version of the micro statement (7). Also important, in understanding (1), is its relationship with the fundamental theorem of the integral calculus, which we state [for comparison with (1)] in this form:

$$\int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b}, \tag{9}$$

if $F'(x)$ is continuous on $a \leq x \leq b$. That is, if the integrand is a derivative and is continuous then the integral can be reduced to an evaluation at the boundary of the integration region, namely, the endpoints a and b. Analogously, the volume integral in (1) can be reduced to an evaluation at the boundary of the region $\mathcal{V}$ (namely, on its surface $\mathcal{S}$) if the integrand is a kind of derivative, namely, a divergence. In fact, (1) reduces to (9) in the one-dimensional case, where $\mathbf{v} = v_x(x,y,z)\hat{\mathbf{i}} + v_y(x,y,z)\hat{\mathbf{j}} + v_z(x,y,z)\hat{\mathbf{k}}$ is merely $v_x(x)\hat{\mathbf{i}}$. Thus, the three-dimensional generalization of the derivative, $F'(x)$ in (9), is the divergence $\operatorname{div}\mathbf{v}$ in (1).

To evaluate the integrals in (1), we need expressions for dV and dA. Of these, dV is easy since it depends only on the coordinate system being used and we know, once and for all, that

$$dV = \begin{cases} dx\,dy\,dz & \text{(Cartesian)} \\ r\,dr\,d\theta\,dz & \text{(cylindrical)} \\ \rho^2\,|\sin\phi|\,d\rho\,d\phi\,d\theta. & \text{(spherical)} \end{cases} \tag{10}$$

However, dA is harder inasmuch as it is specific to the particular surface under consideration. The general formula for dA was given in Section 15.5 by (4) and (5), and special cases were covered by (13) and (18) therein. Since you may not have studied Section 15.5, let us try to get by, in this section and the next, relying only on equation (18) from Section 15.5, namely,

$$dA = \sqrt{1 + f_x^2 + f_y^2}\,dx\,dy \tag{11}$$

if the surface $\mathcal{S}$ is known in the form $z = f(x,y)$. The subscripts in (11) denote partial derivatives. Here, dA is the desired area element on $\mathcal{S}$ and $dx\,dy$ is the projection, or shadow, of that area onto the x, y plane (Fig. 3).

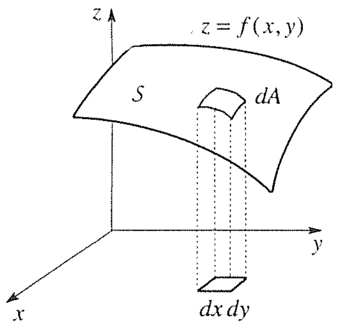


Figure 3. Interpretation of (11).

In the example to follow, we illustrate the "mechanics" of the calculations in (1), and then we turn to physical applications.

EXAMPLE 1. *Verification of (1).* Let $\mathbf{v} = x(y+1)z^3\hat{\mathbf{j}}$, and let $\mathcal{V}$ be the pentahedron shown in Fig. 4, with faces given by the planes $x = 0$, $x = 2$, $y = 0$, $z = 0$, and the slanted plane $EFGH$. The problem that we pose is merely to evaluate both sides of (1), for that case, and to verify that they are equal. Naturally, such verification by no means proves the truth of (1), but if the left- and right-hand sides of (1) are unequal, for this single example, that result would suffice to *dis*prove (1).

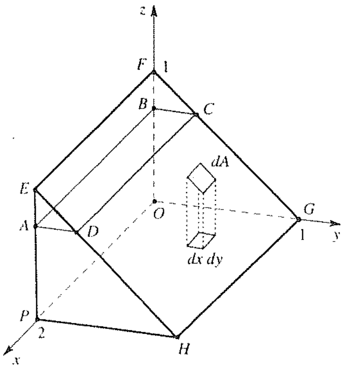


Figure 4. The pentahedron $\mathcal{V}$.

We will need the equation of the plane $EFGH$. We know it is of the form

$$ax + by + cz = d, \tag{12}$$

and if we force the plane to go through three noncollinear point, such as E, F, and G, that should uniquely determine its equation. Since $E = (2,0,1)$, $F = (0,0,1)$, and $G = (0,1,0)$, we obtain, from (12), $2a + c = d$, $c = d$, and $b = d$. Thus, $a = 0$, $b = d$, $c = d$ so (12) becomes $dy + dz = d$, or

$$y + z = 1. \tag{13}$$

We begin with the left-hand side of (1):

$$\begin{aligned}\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV &= \int_{\mathcal{V}} \left\{ \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}[x(y+1)z^3] + \frac{\partial}{\partial z}(0) \right\} dV \\ &= \int_0^1 \left\{ \int_0^{1-z} \int_0^2 xz^3\, dx\, dy \right\} dz \\ &= 2 \int_0^1 z^3(1-z)\, dz = \frac{1}{10}, \end{aligned} \tag{14}$$

where the limits $\int_0^{1-z}\int_0^2$ correspond to the shaded rectangle $ABCD$.

Turning to the surface integral in (1), we consider the five faces separately. Since those faces are planes, $\hat{\mathbf{n}}$ is simply a constant over each one. On the $x = 0$ face $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$, and $\hat{\mathbf{n}} \cdot \mathbf{v} = -\hat{\mathbf{i}} \cdot x(y+1)z^3\hat{\mathbf{j}} = 0$, so $\int \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = \int 0\, dA = 0$ over that face. Similarly, for the $z = 0$ face, since $\hat{\mathbf{n}} \cdot \mathbf{v} = (-\hat{\mathbf{k}}) \cdot x(y+1)z^3\hat{\mathbf{j}} = 0$ there. On $y = 0$, dA is $dx\, dz$, and

$$\begin{aligned}\int_{y=0 \text{ face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{y=0 \text{ face}} -\hat{\mathbf{j}} \cdot [x(y+1)z^3\hat{\mathbf{j}}]\Big|_{y=0} dA \\ &= -\int_0^1 \int_0^2 xz^3\, dx\, dz = -\frac{1}{2}. \end{aligned} \tag{15}$$

Finally, consider the slanted face. We need expressions for $\hat{\mathbf{n}}$ and dA. One way to compute $\hat{\mathbf{n}}$ is to cross the vector $\mathbf{HG}$ (from H to G) into the vector $\mathbf{HE}$ (from H to E):

$$\mathbf{n} = \mathbf{HG} \times \mathbf{HE} = -2\hat{\mathbf{i}} \times (-\hat{\mathbf{j}} + \hat{\mathbf{k}}) = 2\hat{\mathbf{k}} + 2\hat{\mathbf{j}}.$$

Normalizing the latter (i.e., to unit length),

$$\hat{\mathbf{n}} = \frac{\mathbf{n}}{\|\mathbf{n}\|} = \frac{2\hat{\mathbf{j}} + 2\hat{\mathbf{k}}}{\sqrt{8}} = \frac{1}{\sqrt{2}}(\hat{\mathbf{j}} + \hat{\mathbf{k}}). \tag{16}$$

Next, (13) gives $z = 1 - y$ so $f(x, y)$ is $1 - y$, $f_x = 0$, $f_y = -1$, and (11) gives

$$dA = \sqrt{1 + (0)^2 + (-1)^2}\, dx\, dy = \sqrt{2}\, dx\, dy. \tag{17}$$

Thus,

$$\begin{aligned}\int_{y+z=1 \text{ face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_0^1 \int_0^2 \left(\frac{\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{2}} \right) \cdot [x(y+1)z^3\hat{\mathbf{j}}]\Big|_{z=1-y} (\sqrt{2}\, dx\, dy) \\ &= \int_0^1 \int_0^2 x(y+1)(1-y)^3\, dx\, dy = \frac{3}{5}. \end{aligned} \tag{18}$$

Adding the results for the five parts of S,

$$\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = 0 + 0 + 0 - \frac{1}{2} + \frac{3}{5} = \frac{1}{10},$$

which does indeed agree with (14).

COMMENT 1. Surely v is C^1 in the closed region $\mathcal{V}$, as required by the divergence theorem since the partial derivatives of $v_x = 0$, $v_y = x(y+1)z^3$, $v_z = 0$ with respect to x, y, z are continuous functions of x, y, z in $\mathcal{V}$.

COMMENT 2. Observe that the formula (17) converts the surface integration on the slanted plane to an equivalent integration over its "shadow" (the rectangle $PHGO$) in the x, y plane; hence the limits $\int_0^1 \int_0^2$ in (18). Also, remember to express any z's in the integrand in terms of x, y according to $z = f(x, y) = 1 - y$. ■

Our cross product method of deriving $\hat{\mathbf{n}}$, in Example 1, works only if the surface is planar. A more general method, for deriving normals to surfaces, is as follows. Suppose the surface is given by the relation

$$g(x, y, z) = c, \tag{19}$$

where c is a constant. If we *imagine* $g(x, y, z)$ to be a scalar field defined in 3-space, say a temperature field for definiteness, then the particular surface $\mathcal{S}$ defined by (19) is an isothermal surface, and ∇g will be normal to that surface. Thus, if we normalize ∇g, then we have the unit normal vector

$$\boxed{\hat{\mathbf{n}} = \pm\frac{\nabla g}{\|\nabla g\|},} \tag{20}$$

where the correct sign is to be chosen so that $\hat{\mathbf{n}}$ is outward, not inward. For instance, in Example 1 the equation of the slanted face is $y + z = 1$, so $g(x, y, z) = y + z$ and

$$\hat{\mathbf{n}} = \pm\frac{\nabla(y+z)}{\|\nabla(y+z)\|} = \pm\frac{\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{2}}. \tag{21}$$

Clearly, the plus sign gives the *out*ward unit normal, as desired, and then (21) agrees with (16), which was found by other means. We will call (20) the **gradient method** for obtaining the field of normals to a surface defined by (19).

Consider two important physical applications.

EXAMPLE 2. *Continuity Equation of Fluid Mechanics.* Consider a fluid flow within some region $\mathcal{R}$ in 3-space; let $\mathbf{v}(x, y, z, t)$ and $\sigma(x, y, z, t)$ be the velocity and density (i.e., the mass per unit volume) fields, respectively.[*] Let $\mathcal{V}$ be a stationary control volume of arbitrary shape (although we ask that its surface be piecewise smooth and orientable), located within $\mathcal{R}$. The mass within $\mathcal{V}$ at any time t is $M = \int_{\mathcal{V}} \sigma\, dV$, so the rate of increase of M is[†]

$$\frac{dM}{dt} = \frac{d}{dt}\int_{\mathcal{V}} \sigma(x, y, z, t)\, dV = \int_{\mathcal{V}} \frac{\partial \sigma}{\partial t}\, dV. \tag{22}$$

[*] The letter ρ is usually used for the mass density, but we will reserve ρ for the ρ in ρ, ϕ, θ spherical coordinate systems.

[†] For the last equality in (22), see equation (12.3) in Exercise 12 of Section 13.8.

Next, observe that the rate at which mass enters $\mathcal{V}$ through its boundary $\mathcal{S}$ is the flux integral (Fig. 5)

$$-\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA, \tag{23}$$

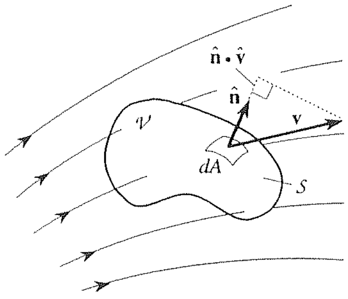


Figure 5. Control volume $\mathcal{V}$.

where $\hat{\mathbf{n}}$ is the unit outward normal to $\mathcal{S}$. [The latter expression differs in sign from (2) in Section 16.3 because there we considered outflow, whereas here we consider inflow. We could avoid the minus sign by taking $\hat{\mathbf{n}}$ to be the *in*ward normal, but it is tradition, in vector field theory, to work with outward normals.]

Finally, suppose that there is a creation of mass within the field at a rate $f(x, y, z, t)$ mass per unit volume per unit time; where f is positive mass is being created, and where f is negative mass is being destroyed. Then the net rate of creation of mass within $\mathcal{V}$ is

$$\int_{\mathcal{V}} f(x, y, z, t)\, dV \tag{24}$$

mass per unit time.

Since the rate of increase of mass within $\mathcal{V}$ is certainly equal to the rate at which mass enters $\mathcal{V}$ through $\mathcal{S}$, plus the rate at which mass is created, it follows from (22)–(24) that

$$\int_{\mathcal{V}} \frac{\partial \sigma}{\partial t}\, dV = -\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\mathcal{V}} f\, dV. \tag{25}$$

To combine these integrals we convert the surface integral to a volume integral using the divergence theorem (1), with "v" in (1) taken to be $\sigma\mathbf{v}$ in (25):

$$\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (\sigma\mathbf{v})\, dA = \int_{\mathcal{V}} \nabla \cdot (\sigma\mathbf{v})\, dV. \tag{26}$$

Thus (25) becomes

$$\int_{\mathcal{V}} \left[\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) - f \right] dV = 0. \tag{27}$$

We could accept (27) as our final result, but its integral form is inconvenient. To achieve the final simplification, observe that the control volume $\mathcal{V}$ is arbitrary. If $\int_{\mathcal{V}} [\ ]\, dV = 0$ for *arbitrary* $\mathcal{V}$, within $\mathcal{R}$, and we are willing to assume that the $[\ ]$ integrand is continuous, then it can be concluded that $[\ ] = 0$ throughout $\mathcal{R}$. For suppose that $[\ ]$ were *non*zero, say positive, at some point P in $\mathcal{R}$. Then by the assumed continuity of $[\ ]$ there must be some neighborhood $\mathcal{N}$ of P such that $[\ ] > 0$ throughout $\mathcal{N}$. Since $\mathcal{V}$ is arbitrary, we can take $\mathcal{V}$ to be $\mathcal{N}$. Then $\int_{\mathcal{V}} [\ ]\, dV = \int_{\mathcal{N}} [\ ]\, dV > 0$, which contradicts (27). Thus, $[\ ]$ must be identically zero through $\mathcal{R}$: $\partial\sigma/\partial t + \nabla \cdot (\sigma\mathbf{v}) - f = 0$, or

$$\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) = f. \tag{28}$$

Thus far, (28) involves logic and bookkeeping, no physics, for who can argue with the statement that the rate of increase of mass within $\mathcal{V}$ equals the rate at which it enters through $\mathcal{S}$ plus the rate at which it is generated within $\mathcal{V}$? Finally, the *physics*: if matter can be neither created nor destroyed, then the mass generation term, f, must be zero, and (28) reduces to the form

$$\boxed{\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) = 0,} \tag{29}$$

which is the well known **continuity equation** of fluid mechanics and is a statement of the physical principle of conservation of mass. It is a partial differential equation, or *field equation*, governing the scalar density field σ and the vector velocity field $\mathbf{v}$. If $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$, then (29) becomes

$$\frac{\partial \sigma}{\partial t} + \frac{\partial}{\partial x}(\sigma v_x) + \frac{\partial}{\partial y}(\sigma v_y) + \frac{\partial}{\partial z}(\sigma v_z) = 0. \tag{30}$$

The latter is only one equation on the four unknowns σ, v_x, v_y, v_z, but (30) is not the only relevant field equation. For instance, Newton's vector second law of motion leads to another field equation relating σ and $\mathbf{v}$, which is a vector equation or, equivalently, three scalar equations. That step would appear to "close" the system since, together with (30), it would give four equations on the four unknowns. However, it also introduces another unknown field, the pressure field $p(x, y, z, t)$ so we need to add to our system of equations an *equation of state*, relating the pressure, density, and temperature fields. Having thus included the temperature field, say $T(x, y, z, t)$, we need to use the law(s) of thermodynamics to obtain the field equation(s) needed to close the system. The upshot is that the various unknown fields, $\sigma, v_x, v_y, v_z, p, T$, are governed by a system of coupled field equations, most of them being PDE's (partial differential equations).

COMMENT. It is often possible to simplify this complex state of affairs. For instance, if the fluid is water then we might very well be able to neglect its compressibility and, to a good approximation, assume that $\sigma =$ constant. In that case $\partial\sigma/\partial t = 0$, so (29) becomes $0 + \nabla \cdot (\sigma\mathbf{v}) = \sigma\nabla \cdot \mathbf{v} = 0$ or, equivalently,

$$\nabla \cdot \mathbf{v} = 0. \tag{31}$$

Then, even though Newton's vector equation of motion introduces the unknown pressure field p, those three equations plus (31) comprise a closed system of four PDE's governing the four unknowns v_x, v_y, v_z, and p. We can still solve the relevant field equation for the temperature field T if we wish, but the point is that if $\sigma =$ constant then that equation stands alone, and is uncoupled from the PDE's governing the velocity and pressure fields.

How do we know whether $\sigma =$ constant is a sufficiently accurate approximation? There is always a relevant nondimensional parameter that tells us whether or not one effect is negligible compared to others. The nondimensional parameter that provides a measure of the effects of compressibility is the *Mach number* M, defined as the ratio of the fluid velocity to the local speed of sound. If $M \ll 1$ we can, with good accuracy, make the simplifying assumption that $\sigma =$ constant, that is, that the flow is incompressible. For instance, since the speed of sound in water is around 1500 m/s, the Mach number within a water wave is, no doubt, tiny compared to unity, so in developing the field equations governing water waves one would surely take the water to be incompressible. ▮

EXAMPLE 3. *Unsteady Heat Conduction.* Consider the unsteady conduction of heat within some region $\mathcal{R}$ in 3-space. Our derivation of a partial differential equation governing the temperature field $T(x, y, z, t)$ will closely parallel our derivation of the continuity equation in the preceding example.

First, we choose a fixed arbitrary control volume $\mathcal{V}$ within $\mathcal{R}$. Instead of a mass balance, this time we carry out a heat balance. That is, we equate the rate at which heat accumulates in $\mathcal{V}$ to the rate at which it enters through the surface $\mathcal{S}$ of $\mathcal{V}$ plus the rate at which it is generated within $\mathcal{V}$.

Since the heat (in calories, say) contained in a mass m at (absolute) temperature T is mcT, where c is the *specific heat* of the material, the rate of accumulation of heat within $\mathcal{V}$ is

$$\frac{d}{dt}\int_{\mathcal{V}} cT\sigma\, dV \qquad \text{or} \qquad \int_{\mathcal{V}} c\sigma \frac{\partial T}{\partial t}\, dV. \tag{32}$$

As in Example 2, σ is the mass density (mass per unit volume) of the material and dV is an infinitesimal volume element. Thus, $\sigma\, dV$ in (32) is $d(\text{mass})$.

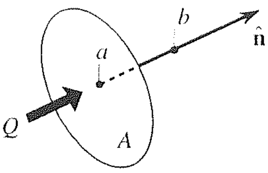


Figure 6. Fourier's law.

To compute the heat flux into $\mathcal{V}$ through $\mathcal{S}$ we need the **Fourier law of heat conduction**, which states that the heat flux Q (calories per second, say) through an area element A is proportional to A and to the derivative of the temperature field normal to the element A (Fig. 6):

$$Q = -kA\frac{\partial T}{\partial n}, \tag{33}$$

where the constant of proportionality, k, is called the thermal *conductivity* of the given material; for instance, k for copper is much larger than k for wood. Why the minus sign? If Q really is positive (i.e., in the $+\hat{\mathbf{n}}$ direction), then $T(a) > T(b)$ (since heat flows from hot to cold) and $\partial T/\partial n$ will be negative.

According to the Fourier law (33), the heat flux into $\mathcal{V}$ through an area element dA on the surface $\mathcal{S}$ of $\mathcal{V}$ is $k\,\frac{\partial T}{\partial n}\, dA$, where there is no minus sign this time because $\hat{\mathbf{n}}$ is the outward unit normal whereas we seek the inward heat flux. Thus, the heat flux into $\mathcal{V}$, through $\mathcal{S}$, is

$$\int_{\mathcal{S}} k\frac{\partial T}{\partial n}\, dA. \tag{34}$$

Further, if heat is being generated within $\mathcal{V}$ at the rate $f(x, y, z, t)$ calories per unit volume per unit time, then the net rate of generation of heat within $\mathcal{V}$ is

$$\int_{\mathcal{V}} f(x, y, z, t)\, dV. \tag{35}$$

And since the rate of increase of heat within $\mathcal{V}$ equals the rate of heat entering through $\mathcal{S}$ plus the rate of generation within, we have

$$\int_{\mathcal{V}} c\sigma\frac{\partial T}{\partial t}\, dV = \int_{\mathcal{S}} k\frac{\partial T}{\partial n}\, dA + \int_{\mathcal{V}} f\, dV. \tag{36}$$

To convert the surface integral in (36) to a volume integral using the divergence theorem $\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\mathbf{v}\, dV = \int_{\mathcal{V}} \nabla\cdot\mathbf{v}\, dV$, we need to identify the vector field "v" such that $k\,\partial T/\partial n = \hat{\mathbf{n}}\cdot\mathbf{v}$. From the directional derivative formula (7) in Section 16.4, with $\hat{\mathbf{s}} = \hat{\mathbf{n}}$, $\partial T/\partial n = \nabla T\cdot\hat{\mathbf{n}}$, so $k\,\partial T/\partial n = k\,\nabla T\cdot\hat{\mathbf{n}}$. Thus "v" $= k\,\nabla T$. Then the divergence theorem gives $\int_{\mathcal{S}} k\,(\partial T/\partial n)\, dA = \int_{\mathcal{S}} k\,\nabla T\cdot\hat{\mathbf{n}}\, dA = \int_{\mathcal{V}} \nabla\cdot(k\,\nabla T)\, dV$ so (36) becomes

$$\int_{\mathcal{V}} \left[c\sigma\frac{\partial T}{\partial t} - \nabla\cdot(k\,\nabla T) - f\right] dV = 0. \tag{37}$$

From the arbitrariness of $\mathcal{V}$ there follows (as discussed in Example 2) the PDE, or field equation,

$$c\sigma\frac{\partial T}{\partial t} - \nabla\cdot(k\,\nabla T) - f = 0 \tag{38}$$

governing the temperature field $T(x, y, z, t)$. If the conductivity k is a constant, then $\nabla \cdot (k\,\nabla T) = k\nabla \cdot \nabla T = k\nabla^2 T$, and (38) may be written as

$$\alpha^2 \nabla^2 T - \frac{\partial T}{\partial t} = -F, \tag{39}$$

where $F = f(x, y, z, t)/(c\sigma)$, and $\alpha^2 = k/(c\sigma)$ is called the *diffusivity* of the material. The $-F$ term renders the PDE (39) nonhomogeneous and serves as a forcing function just as the $F_0 \sin \Omega t$ term in the ODE $mx'' + kx = F_0 \sin \Omega t$ makes that equation nonhomogeneous and serves as a forcing function. If $F(x, y, z, t) \equiv 0$, one usually writes (39) in the form

$$\boxed{\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t},} \tag{40}$$

which is called the **heat equation** governing the unsteady diffusion of heat by conduction, in a material having diffusivity α^2. If a steady state is achieved, then $\partial T/\partial t = 0$ and (40) reduces to the **Laplace equation**

$$\boxed{\nabla^2 T = 0.} \tag{41}$$

COMMENT 1. In applications, the source term F in (39) may be nonzero. For example, there might be taking place within the region a chemical reaction that releases heat (exothermic) or absorbs heat (endothermic). Or, a mechanical member (such as a beam) subjected to a cyclic loading might be heated by mechanical hysteresis losses proportional to the local stress level.

COMMENT 2. The conduction of heat, in a medium, is an example of the physical process of diffusion, so (40) is also known as the **diffusion equation**. Another example is the diffusion of material, such as the spreading out by diffusion of a drop of dye place in a pot of water. Let $c(x, y, z, t)$ denote the concentration of the material (i.e., the mass of material per unit volume of medium). Analogous to the Fourier law (33) is **Fick's law**, that the material mass flux through an area element A is proportional to A and to the derivative $\partial c/\partial n$ normal to A. Consequently, just like the temperature field T, the concentration field c satisfies a diffusion equation

$$\beta^2 \nabla^2 c = \frac{\partial c}{\partial t}, \tag{42}$$

where β^2 is a mass diffisivity constant. Arguably, the diffusion equation and the Laplace equation are two of the three most prominent PDE's of mathematical physics, the third being the wave equation, which we will study in Chapter 20. ▮

EXAMPLE 4. *Green's Identities.* If u and v are scalar fields, then the combination $u\nabla v$ is a vector field. Applying the divergence theorem to that field gives

$$\int_{\mathcal{V}} \nabla \cdot (u\nabla v)\, dV = \int_{S} \hat{\mathbf{n}} \cdot (u\nabla v)\, dA. \tag{43}$$

But from equation (4) in Section 16.6, with ∇v as "v," we have $\nabla \cdot (u\nabla v) = \nabla u \cdot \nabla v + u\nabla \cdot \nabla v = \nabla u \cdot \nabla v + u\nabla^2 v$, and $\hat{\mathbf{n}} \cdot (u\nabla v) = u\,\hat{\mathbf{n}} \cdot \nabla v = u\,\partial v/\partial n$, so

$$\boxed{\int_{\mathcal{V}} (\nabla u \cdot \nabla v + u\nabla^2 v)\, dV = \int_{S} u\, \frac{\partial v}{\partial n}\, dA.} \tag{44}$$

This formula, known as **Green's first identity**,* is important in the study of partial differential equations.

In some cases, evaluation of the $\partial v/\partial n$ term in (44) is simple. For example, if V is the cube $|x| < 1$, $|y| < 1$, $|z| < 1$, then $\partial v/\partial n$ on the $x = +1$ face is merely $\partial v/\partial x$, and on the $x = -1$ face it is $-\partial v/\partial x$. For more complicated surface shapes we suggest that you use the expression $\partial v/\partial n = \nabla v \cdot \hat{\mathbf{n}}$ to calculate $\partial v/\partial n$ on $\mathcal{S}$.

Now, interchanging the letters u and v in (44), it should be equally true that

$$\int_{\mathcal{V}} (\nabla v \cdot \nabla u + v\nabla^2 u)\, dV = \int_{\mathcal{S}} v \frac{\partial u}{\partial n}\, dA,$$

and subtracting these last two equations yields **Green's second identity**

$$\boxed{\int_{\mathcal{V}} (u\nabla^2 v - v\nabla^2 u)\, dV = \int_{\mathcal{S}} \left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n} \right) dA.} \tag{45}$$

What conditions on u, v assure the validity of Green's identities? Recall that the three components of $\mathbf{v}$ in the divergence theorem (1) were to have continuous first-order partials with respect to x, y, z. Since in deriving (44) and (45) we first set "$\mathbf{v}$" $= u\nabla v = uv_x\hat{\mathbf{i}} + uv_y\hat{\mathbf{j}} + uv_z\hat{\mathbf{k}}$, and then "$\mathbf{v}$" $= v\nabla u = vu_x\hat{\mathbf{i}} + vu_y\hat{\mathbf{j}} + vu_z\hat{\mathbf{k}}$, we must ask uv_x, uv_y, uv_z, vu_x, vu_y, and vu_z to have continuous first-order partial derivatives in the closed region $\mathcal{V}$. Equivalently, we ask u and v to be C^2 in the closed region $\mathcal{V}$. Similarly, $\mathcal{S}$ is to be a piecewise smooth orientable surface, as in the divergence theorem. ▮

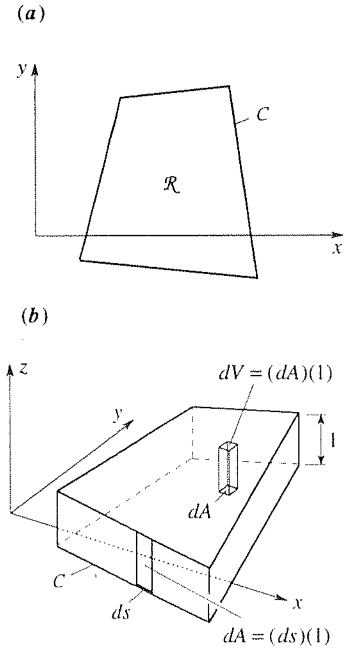


Figure 7. Plane case.

16.8.2. Two-dimensional case. Suppose $\mathbf{v}$ is but a *two*-dimensional field, $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$, defined over a closed region $\mathcal{R}$ in the x, y plane (Fig. 7a). To have a three-dimensional region, so we can apply the divergence theorem (1), let us build up a volume $\mathcal{V}$ with $\mathcal{R}$ as its base, and of unit thickness in the z direction, as in Fig. 7b. In the divergence theorem

$$\begin{aligned}\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV &= \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA \\ &= \int_{\text{top}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\text{bottom}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\text{side}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA,\end{aligned} \tag{46}$$

$\mathcal{S}$ is comprised of a top ($z = 1$) on which $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$, a bottom ($z = 0$) on which $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$, and a side on which $\hat{\mathbf{n}}$ is parallel to the x, y plane. Since $\hat{\mathbf{n}} \cdot \mathbf{v} = 0$ on the top and bottom, the "top" and "bottom" integrals in (46) are zero. Further, $\operatorname{div} \mathbf{v}$ in (46) does not vary with z, so we can let dV extend from bottom to top; then $dV = (dA)(1)$, where dA is the base area of dV (Fig. 7b). Similarly, $\hat{\mathbf{n}} \cdot \mathbf{v}$ on the side face does not vary with z, so we can let dA in the side integral extend from

***George Green* (1793–1841) was a self-taught English mathematical physicist. His work on electricity and magnetism helped to place field theory on a firm mathematical foundation. His "Green's identities," "Green's theorem," and "Green's functions" are important tools in the study of field theory and partial differential equations.

bottom to top (shaded in Fig. 7b); then $dA = (ds)(1)$, where ds is the arc length along $\mathcal{C}$. With these replacements, (46) becomes

$$\boxed{\int_{\mathcal{R}} \nabla \cdot \mathbf{v}\, dA = \int_{\mathcal{C}} \hat{\mathbf{n}} \cdot \mathbf{v}\, ds,} \tag{47}$$

which we call the **two-dimensional divergence theorem**. That is, we simply drop back by one dimension: $\mathcal{V} \to \mathcal{R}$, $dV \to dA$, $\mathcal{S} \to \mathcal{C}$, and $dA \to ds$, The integral on the right-hand side of (47) is a line integral. Line integrals were defined in Section 15.2.3.

EXAMPLE 5. Let us illustrate (47) with a simple example. Namely, verify (47) for the case where $\mathbf{v} = x^2 y\hat{\mathbf{i}} - 3\hat{\mathbf{j}}$ and $\mathcal{R}$ is the rectangle shown in Fig. 8. Then

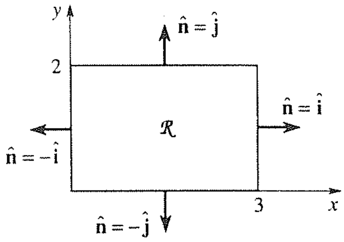


Figure 8. The region $\mathcal{R}$.

$$\int_{\mathcal{R}} \nabla \cdot \mathbf{v}\, dA = \int_0^2 \int_0^3 2xy\, dx\, dy = 18,$$

and

$$\begin{aligned}\int_{\mathcal{C}} \hat{\mathbf{n}} \cdot \mathbf{v}\, ds &= \int_0^3 (-\hat{\mathbf{j}}) \cdot \mathbf{v}\Big|_{y=0} dx + \int_0^2 \hat{\mathbf{i}} \cdot \mathbf{v}\Big|_{x=3} dy \\ &\quad + \int_0^3 \hat{\mathbf{j}} \cdot \mathbf{v}\Big|_{y=2} dx + \int_0^2 (-\hat{\mathbf{i}}) \cdot \mathbf{v}\Big|_{x=0} dy \\ &= \int_0^3 3\, dx + \int_0^2 9y\, dy - \int_0^3 3\, dx + \int_0^2 0\, dy = 18,\end{aligned}$$

which results are indeed the same.

COMMENT. In the integral $\int_0^2 (-\hat{\mathbf{i}}) \cdot \mathbf{v}\Big|_{x=0} dy$, say, how do we know that the limits are 0 to 2, not 2 to 0? The idea is that the ds in (47) came from "$(ds)(1)$," namely, the area element dA, and since area elements are necessarily positive, we need to choose integration limits so that the ds increments are all positive. Hence we let y go from 0 to 2, not from 2 to 0. Similarly for the integrals on the edges $x = 3$, $y = 0$, and $y = 2$. ▮

16.8.3. Non-Cartesian coordinates. (Optional) Let us work two examples involving cylindrical and spherical coordinates.

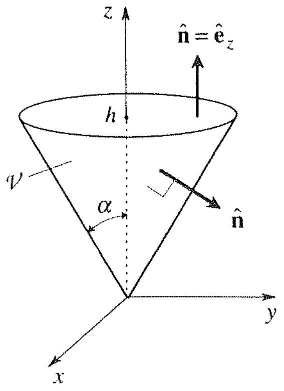


Figure 9. The cone $\mathcal{V}$.

EXAMPLE 6. *Cylindrical Coordinates.* Let $\mathbf{v} = r^2 z\hat{\mathbf{e}}_z$, and let $\mathcal{V}$ be the cone shown in Fig. 9. The problem that we pose is to verify the divergence theorem, for that case, by working out the left- and right-hand sides of (1) and verifying that they are equal. Insofar as $\mathcal{V}$ is concerned, cylindrical and spherical coordinates are equally convenient: in cylindrical coordinates the flat top is a constant-coordinate surface ($z = h$) but the conical part is not, and in spherical coordinates the conical part is a constant-coordinate surface ($\phi = \alpha$) but the flat top is not. However, $\mathbf{v}$ is given in cylindrical coordinates, so let us use cylindrical coordinates r, θ, z.

Equation (16) in Section 16.7 gives

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}[r(0)] + \frac{1}{r}\frac{\partial}{\partial \theta}(0) + \frac{\partial}{\partial z}(r^2 z) = r^2,$$

and the volume element is $dV = r\,dr\,d\theta\,dz$, so

$$\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\,dV = \int_0^h \int_0^{2\pi} \int_0^{z\tan\alpha} (r^2)\,r\,dr\,d\theta\,dz = \frac{\pi}{10}h^5 \tan^4\alpha, \tag{48}$$

where the r limits follow from the fact that the equation of the conical surface is $r = (\tan\alpha)z$.

To evaluate the right-hand side of (1) we break the integral into an integral over the flat top plus an integral on the conical surface. On the top we have $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ and $dA = r\,dr\,d\theta$, so*

$$\begin{aligned}\int_{\text{top}} \hat{\mathbf{n}} \cdot \mathbf{v}\,dA &= \int_0^{2\pi} \int_0^{h\tan\alpha} (r^2 z)\Big|_{z=h} r\,dr\,d\theta \\ &= h\int_0^{2\pi} \int_0^{h\tan\alpha} r^3\,dr\,d\theta = \frac{\pi}{2}h^5 \tan^4\alpha.\end{aligned} \tag{49}$$

To determine $\hat{\mathbf{n}}$ on the conical surface we use the gradient method explained in connection with equations (19) and (20). That is, since the surface is defined by $r = (\tan\alpha)z$, we have

$$\begin{aligned} g(r,\theta,z) &= r - (\tan\alpha)z = 0, \\ \mathbf{n} = \pm\nabla g &= \pm\left(\hat{\mathbf{e}}_r \frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta \frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\mathbf{e}}_z \frac{\partial}{\partial z}\right)[r - (\tan\alpha)z] \\ &= \pm\left[\hat{\mathbf{e}}_r - (\tan\alpha)\hat{\mathbf{e}}_z\right], \\ \|\mathbf{n}\| &= \sqrt{1+\tan^2\alpha} = \frac{1}{\cos\alpha}, \\ \hat{\mathbf{n}} &= +\frac{\hat{\mathbf{e}}_r - (\tan\alpha)\hat{\mathbf{e}}_z}{1/\cos\alpha} = (\cos\alpha)\hat{\mathbf{e}}_r - (\sin\alpha)\hat{\mathbf{e}}_z, \end{aligned} \tag{50}$$

where the plus sign was selected since it gives the *out*ward normal. Next, we need the area element dA on the conical surface. For that we rely on the general formula (5) in Section 15.5, $dA = \sqrt{EG - F^2}\,du\,dv$. To use that formula we need parametric equations $x = x(u,v)$, $y = y(u,v)$, $z = z(u,v)$ for the surface, which can be obtained by beginning with the equations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$ relating x, y, z to r, θ, z, and adding the fact that the cone is given by $z = (\cot\alpha)r$:

$$\begin{aligned} x &= r\cos\theta = u\cos v, \\ y &= r\sin\theta = u\sin v, \\ z &= (\cot\alpha)r = (\cot\alpha)u. \end{aligned} \tag{51}$$

That is, let r and θ be u and v, respectively.

$$\begin{aligned} E &= x_u^2 + y_u^2 + z_u^2 = \cos^2 v + \sin^2 v + \cot^2\alpha = \sec^2\alpha, \\ F &= x_u x_v + y_u y_v + z_u z_v = (\cos v)(-u\sin v) + (\sin v)(u\cos v) = 0, \\ G &= x_v^2 + y_v^2 + z_v^2 = u^2\sin^2 v + u^2\cos^2 v = u^2 \end{aligned}$$

*If you don't see why $dA = r\,dr\,d\theta$, see (17) in Section 15.6.

so

$$dA = \sqrt{EG - F^2}\, du\, dv = (\operatorname{cosec} \alpha)\, u\, du\, dv. \tag{52}$$

Thus, remembering that $r = u$ and $z = (\cot \alpha)\, u$ on the cone,

$$\begin{aligned}\int_{\text{cone}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{\text{cone}} [(\cos \alpha)\hat{\mathbf{e}}_r - (\sin \alpha)\hat{\mathbf{e}}_z] \cdot (r^2 z \hat{\mathbf{e}}_z)(\operatorname{cosec} \alpha) u\, du\, dv \\ &= -\int_{\text{cone}} r^2 z u\, du\, dv \\ &= -\int_0^{2\pi} \int_0^{h \tan \alpha} (\cot \alpha) u^4\, du\, dv \\ &= -\frac{2\pi}{5} h^5 \tan^4 \alpha. \end{aligned} \tag{53}$$

Adding (49) and (53) we obtain

$$\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = \frac{\pi}{10} h^5 \tan^4 \alpha,$$

which does agree with (48).

COMMENT 1. Though it is clear by inspection that $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the flat top, observe that the gradient method could have been used instead, for the equation of that surface is $z = h$, so $g(r, \theta, z) = z$ and

$$\hat{\mathbf{n}} = \pm \frac{\nabla g}{\|\nabla g\|} = \pm \hat{\mathbf{e}}_z \to +\hat{\mathbf{e}}_z.$$

COMMENT 2. To check results, we urge you to consider special cases. For instance, to check the result $\hat{\mathbf{n}} = (\cos \alpha)\hat{\mathbf{e}}_r - (\sin \alpha)\hat{\mathbf{e}}_z$, derived in (50), observe that for the special cases $\alpha = 0$ and $\alpha = \pi/2$ we have $\hat{\mathbf{n}} = \hat{\mathbf{e}}_r$ and $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_z$, respectively, which results are obviously correct.

COMMENT 3. To see that $\mathbf{v}$ is C^1 in $\mathcal{V}$, as called for by the divergence theorem, express $\mathbf{v} = r^2 z \hat{\mathbf{e}}_z = (x^2 + y^2) z \hat{\mathbf{k}}$, and observe that the partial derivatives of $v_x = 0$, $v_y = 0$, and $v_z = (x^2 + y^2) z$ with respect to x, y, z are continuous functions of x, y, z in $\mathcal{V}$. ▮

EXAMPLE 7. *Gauss's Law.* An electrical charge q (which could be positive or negative) at the origin gives rise to an electric field $\mathbf{E}$ that is expressed most conveniently in terms of spherical coordinates as

$$\mathbf{E} = \frac{q}{4\pi\epsilon} \frac{1}{\rho^2} \hat{\mathbf{e}}_\rho, \tag{54}$$

where ϵ is a physical constant known as the electric permittivity of the medium. That field is radially directed and spherically symmetric, with a magnitude that tends to zero as $\rho \to \infty$ and to infinity as $\rho \to 0$ (since it is proportional to $1/\rho^2$), as hinted at in Fig. 10. A fundamental result in electrostatics, known as **Gauss's law**, states that if S is a smooth closed surface within the medium, then

$$\int_S \hat{\mathbf{n}} \cdot \mathbf{E}\, dA = \begin{cases} 0 & \text{if } S \text{ does not enclose the charge} \\ \dfrac{q}{\epsilon} & \text{if } S \text{ does enclose the charge.} \end{cases} \tag{55}$$

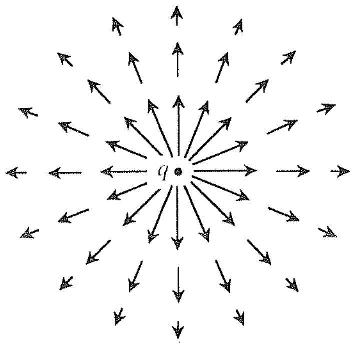


Figure 10. Charge-induced E field.

Our purpose, in this example, is to prove Gauss's law.

First, suppose that $\mathcal{S}$ does not enclose the charge at the origin. Then the divergence theorem gives

$$\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA = \int_{\mathcal{V}} \nabla \cdot \mathbf{E}\, dV = \int_{\mathcal{V}} 0\, dV = 0, \tag{56}$$

where $\mathcal{V}$ is the volume enclosed by $\mathcal{S}$, because [according to (32) in Section 16.7, with "$\mathbf{v}$" $= \mathbf{E}$, "v_ρ" $= q/(4\pi\epsilon\rho^2)$, and "v_ϕ" $=$ "v_θ" $= 0$]

$$\nabla \cdot \mathbf{E} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left(\frac{q}{4\pi\epsilon}\right) + 0 + 0 = 0. \tag{57}$$

However, we cannot claim that (56) holds if $\mathcal{S}$ encloses the origin because the $\mathbf{E}$ field fails to satisfy the conditions of the divergence theorem. Specifically, $\mathbf{E}$ is undefined at the origin because the $1/\rho^2$ is $1/0$ there. Similarly for the first-order partial derivatives of the x, y, z components of

$$\mathbf{E} = \frac{q}{4\pi\epsilon}\frac{1}{\rho^2}\hat{\mathbf{e}}_\rho = \frac{q}{4\pi\epsilon}\frac{1}{x^2+y^2+z^2}\frac{x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}}{\sqrt{x^2+y^2+z^2}},$$

with respect to x, y, z, so $\mathbf{E}$ is not C^1 in $\mathcal{V}$, as required by the divergence theorem.

To overcome this difficulty let us cut the singular point out, so we can stay away from it. That is, let us apply the divergence theorem not to $\mathcal{V}$, but to $\mathcal{V}'$, say, which is the same as $\mathcal{V}$ but with a sphere of radius ρ_0 cut out at the origin (Fig. 11). Then the surface of $\mathcal{V}'$ consists of the two parts $\mathcal{S}$ and $\mathcal{S}_0$; note that the outward (i.e., out of $\mathcal{V}'$) unit normal $\hat{\mathbf{n}}$ on $\mathcal{S}_0$ is $-\hat{\mathbf{e}}_\rho$. Now, the $\mathbf{E}$ field *is* C^1 within $\mathcal{V}'$ (because the origin has been cut out), so the divergence theorem gives

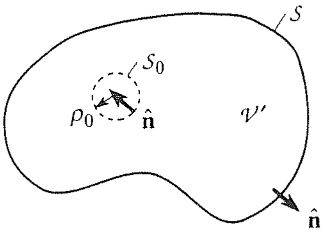


Figure 11. The region $\mathcal{V}'$.

$$\int_{\mathcal{V}'} \nabla \cdot \mathbf{E}\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA + \int_{\mathcal{S}_0} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA \tag{58}$$

or, since $\nabla \cdot \mathbf{E} = 0$ within $\mathcal{V}'$,

$$\begin{aligned}\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA &= -\int_{\mathcal{S}_0} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA \\ &= -\int_{\mathcal{S}_0} -\hat{\mathbf{e}}_\rho \cdot \frac{q}{4\pi\epsilon}\frac{1}{\rho^2}\hat{\mathbf{e}}_\rho \bigg|_{\rho=\rho_0} dA \\ &= \frac{q}{4\pi\epsilon\rho_0^2}\int_{\mathcal{S}_0} dA = \frac{q}{4\pi\epsilon\rho_0^2}4\pi\rho_0^2 = \frac{q}{\epsilon},\end{aligned}$$

as was to be shown. ▮

Closure. The divergence theorem, Theorem 16.8.1, tells us that an integral over a volume $\mathcal{V}$ can be reduced to an integration over the boundary of $\mathcal{V}$, its surface $\mathcal{S}$, provided that its integrand is a divergence (and provided that the surface $\mathcal{S}$ and the vector field $\mathbf{v}$ are sufficiently "decent"). We note that (1) is actually a generalization of the familiar result, from the integral calculus, that the integral $\int_a^b F(x)\, dx$ can be reduced to an evaluation at the boundary of the x interval, its endpoints $x = a$ and $x = b$, provided that $f(x)$ is a derivative, for then

$$\int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b}. \tag{59}$$

Whereas (59) reduces the calculation from one dimension (the line segment on the x axis) to none (the two endpoints), the divergence theorem (1) reduces the calculation from three dimensions (the region $\mathcal{V}$) to two (its surface $\mathcal{S}$). Of course we don't have to use (1) in that direction. In Example 2 and 3, for example, we used it to convert surface integrals to volume integrals.

Our examples are of two types. One type involves the verification of (1) for a particular case, that is, for a particular field $\mathbf{v}$ and region $\mathcal{V}$, the purpose being to foster understanding of the (1) through having to face up to obtaining expressions for $\nabla \cdot \mathbf{v}$, dV, $\hat{\mathbf{n}}$, and dA, and having to evaluate the two integrals. The other type involves the derivation of important results in fluid mechanics, heat conduction, and electrostatics. We will follow that same pattern in the next section as well.

In Examples 2 and 3 we derive two of the most important partial differential equations of mathematical physics, the heat equation and the Laplace equation. In Chapters 18–20 we return to these PDE's and learn how to solve them. Thus, Chapters 13–16 lay much of the groundwork for our study of PDE's in Chapters 18–20.

EXERCISES 16.8

1. In each case, verify the divergence theorem by working out $\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV$ and $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ and showing that the results are equal.

(a) $\mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}}$, $\mathcal{V}$: the rectangular prism $0 \le x \le 1$, $0 \le y \le 3$, $0 \le z \le 2$
(b) $\mathbf{v} = x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}}$, $\mathcal{V}$: the cube $|x| \le 1$, $|y| \le 1$, $|z| \le 1$
(c) $\mathbf{v} = y\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + x\hat{\mathbf{k}}$, $\mathcal{V}$: the rectangular prism $|x| \le 2$, $0 \le y \le 1$, $0 \le z \le 1$
(d) $\mathbf{v} = \hat{\mathbf{j}} + x^2 z\hat{\mathbf{k}}$, $\mathcal{V}$: the cube $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(e) $\mathbf{v} = 3xy^2\hat{\mathbf{j}}$, $\mathcal{V}$: the rectangular prism $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(f) $\mathbf{v} = x^2\hat{\mathbf{i}} - 2z\hat{\mathbf{j}}$, $\mathcal{V}$: the rectangular prism $|x| \le 2$, $|y| \le 2$, $|z| \le 3$, with the cubical cavity $0 < x < 1$, $0 < y < 1$, $0 < z < 1$
(g) $\mathbf{v} = xy\hat{\mathbf{j}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(0,0,1)$, $(1,0,1)$, $(1,2,0)$
(h) $\mathbf{v} = y^2 z\hat{\mathbf{k}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(2,0,0)$, $(0,0,3)$, $(2,0,3)$, $(0,4,3)$, $(2,4,3)$
(i) $\mathbf{v} = x^2 y \sin z\hat{\mathbf{i}}$, $\mathcal{V}$: same as in part (h)
(j) $\mathbf{v} = (3x^2 - 2yz)\hat{\mathbf{j}}$, $\mathcal{V}$: same as in part (h)
(k) $\mathbf{v} = 6\hat{\mathbf{i}} + 2x\hat{\mathbf{j}} + x^2 yz\hat{\mathbf{k}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(1,2,0)$, $(0,0,5)$, $(0,2,5)$
(l) $\mathbf{v} = z^2\hat{\mathbf{k}}$, $\mathcal{V}$: the tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 0$, $2x + y + 2z = 2$
(m) $\mathbf{v} = x^2 z\hat{\mathbf{i}}$, $\mathcal{V}$: same as in part (l)
(n) $\mathbf{v} = x^2 z\hat{\mathbf{i}} - 2z(x^2 + 1)\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, $\mathcal{V}$: same as in part (l)
(o) $\mathbf{v} = z^2\hat{\mathbf{k}}$, $\mathcal{V}$: the tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 1$, $z = x + y$

2. Let $\mathcal{S}$ be a piecewise smooth orientable closed surface enclosing a region of volume V. Show that

(a) $\int_{\mathcal{S}} \hat{\mathbf{n}}\, dA = \mathbf{0}$ HINT: Show that $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{a}\, dA = 0$ for every constant vector $\mathbf{a}$.
(b) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}})\, dA = V$
(c) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}} + y\hat{\mathbf{j}})\, dA = 2V$
(d) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})\, dA = 3V$

3. Verify Green's first identity (44) in each case.

(a) $u = 2x^3$, $v = xy^2$, $\mathcal{V}$: the rectangular prism $0 \le x \le 2$, $0 \le y \le 3$, $0 \le z \le 1$
(b) $u = 1$, $v = x^2 + y^2 + z^2$, $\mathcal{V}$: the cube $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(c) $u = 1$, $v = x^2+y^2+z^2$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,0,1)$, $(1,1,0)$

4. Verify the two-dimensional divergence theorem (47) in each case.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$, $\mathcal{R}$: the rectangle $0 \le x \le a$, $0 \le y \le b$
(b) $\mathbf{v} = x^2y\hat{\mathbf{i}}$, $\mathcal{R}$: the rectangle $0 \le x \le 2$, $0 \le y \le 1$
(c) $\mathbf{v} = xy^2\hat{\mathbf{i}}$, $\mathcal{R}$: the triangle with vertices at $(0,0)$, $(2,0)$, $(2,1)$
(d) $\mathbf{v} = xy\hat{\mathbf{j}}$, $\mathcal{R}$: the triangle with vertices at $(0,0)$, $(2,0)$, $(0,1)$

5. In each case evaluate $I = \int_0^1 \int_0^1 \int_0^1 f \, dx \, dy \, dz$ directly. Then determine a vector field $\mathbf{v}$ such that $\nabla \cdot \mathbf{v} = f$, convert I to a surface integral by using the divergence theorem, and evaluate the surface integral.

(a) $f = x^2yz$ (b) $f = xz^2 - 2y$
(c) $f = 4$ (d) $f = x + y^2z$

6. Derive the following results from results in this section.

(a) $\displaystyle\int_{\mathcal{V}} \nabla^2 u \, dV = \int_{\mathcal{S}} \frac{\partial u}{\partial n} \, dA$

(b) $\displaystyle\int_{\mathcal{V}} \left[\|\nabla u\|^2 + u\nabla^2 u \right] dV = \int_{\mathcal{S}} u \frac{\partial u}{\partial n} \, dA$

CAUTION: Most authors write $(\nabla u)^2$ for $\nabla u \cdot \nabla u$, in place of our $\|\nabla u\|^2$.

7. Just as (47) is the two-dimensional version of (1), work out the two-dimensional versions of Green's

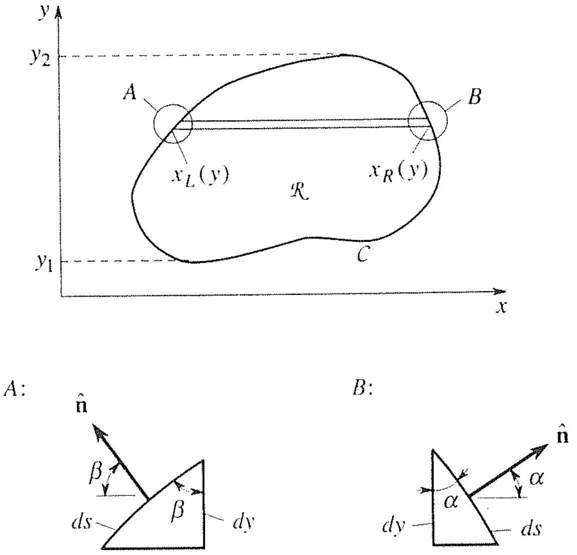
(a) first identity (b) second identity

8. (*Alternative proof of divergence theorem*) An alternative approach to the proof of the divergence theorem (1) is to express the volume integral as an iterated integral, and to carry out one integration. To illustrate the procedure, consider the simpler *two*-dimensional case (47), and suppose that $\mathcal{R}$ is *convex* in the x direction, i.e., each crosshatched sliver running from $x_L(y)$ to $x_R(y)$ (see sketch) lies entirely within $\mathcal{R}$. Now

$$\begin{aligned}\int_{\mathcal{R}} \nabla \cdot \mathbf{v} \, dA &= \int_{\mathcal{R}} \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} \right) dx \, dy \\ &= \int_{\mathcal{R}} \frac{\partial v_x}{\partial x} \, dx \, dy + \int_{\mathcal{R}} \frac{\partial v_y}{\partial y} \, dx \, dy.\end{aligned}$$

The first term becomes

$$\begin{aligned}\int_{\mathcal{R}} \frac{\partial v_x}{\partial x} \, dx \, dy &= \int_{y_1}^{y_2} \int_{x_L(y)}^{x_R(y)} \frac{\partial v_x}{\partial x} \, dx \, dy \\ &= \int_{\text{right}} v_x(ds \cos\alpha) - \int_{\text{left}} v_x(ds \cos\beta) \\ &= \int_{\text{right}} v_x(\hat{\mathbf{i}} \cdot \hat{\mathbf{n}}) \, ds - \int_{\text{left}} v_x(-\hat{\mathbf{i}} \cdot \hat{\mathbf{n}}) \, ds \\ &= \int_{\text{right}} (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}} \, ds + \int_{\text{left}} (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}} \, ds \\ &= \int_C (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}} \, ds.\end{aligned} \tag{8.1}$$



We now state the problem: assuming that $\mathcal{R}$ is convex in the y direction as well, show that

$$\int_{\mathcal{R}} \frac{\partial v_y}{\partial y} \, dx \, dy = \int_C (v_y\hat{\mathbf{j}}) \cdot \hat{\mathbf{n}} \, ds, \tag{8.2}$$

and hence infer (47) from (8.1) and (8.2). NOTE: Extension to nonconvex regions is not difficult but will not be considered.

9. Besides the divergence theorem, $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \hat{\mathbf{v}} \, dA = \int_{\mathcal{V}} \nabla \cdot \hat{\mathbf{v}} \, dV$, one may derive the "companion" results

$$\boxed{\int_{\mathcal{S}} \hat{\mathbf{n}} \, u \, dA = \int_{\mathcal{V}} \nabla u \, dV} \tag{9.1}$$

and

$$\boxed{\int_S \hat{\mathbf{n}} \times \hat{\mathbf{v}}\, dA = \int_V \nabla \times \hat{\mathbf{v}}\, dV.} \tag{9.2}$$

(a) Derive (9.1). HINT: Write

$$\begin{aligned}\int_S \hat{\mathbf{n}}\, u\, dA &= \int_S [(\hat{\mathbf{n}}\cdot\hat{\mathbf{i}})\hat{\mathbf{i}} + (\hat{\mathbf{n}}\cdot\hat{\mathbf{j}})\hat{\mathbf{j}} + (\hat{\mathbf{n}}\cdot\hat{\mathbf{k}})\hat{\mathbf{k}}]u\, dA \\ &= \hat{\mathbf{i}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{i}})\, u\, dA + \hat{\mathbf{j}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{j}})\, u\, dA \\ &\quad + \hat{\mathbf{k}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{k}})\, u\, dA \\ &= \hat{\mathbf{i}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{i}})\, dA + \hat{\mathbf{j}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{j}})\, dA \\ &\quad + \hat{\mathbf{k}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{k}})\, dA,\end{aligned}$$

apply the divergence theorem to each integral, and combine the results into one integral.

(b) Derive (9.2). HINT: Essentially the same hint as in part (a). Write

$$\begin{aligned}\int_S \hat{\mathbf{n}} \times \hat{\mathbf{v}}\, dA &= \int_S \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \hat{\mathbf{n}}\cdot\hat{\mathbf{i}} & \hat{\mathbf{n}}\cdot\hat{\mathbf{j}} & \hat{\mathbf{n}}\cdot\hat{\mathbf{k}} \\ v_x & v_y & v_z \end{vmatrix} dA \\ &= \hat{\mathbf{i}}\int_S \hat{\mathbf{n}}\cdot(v_z\hat{\mathbf{j}} - v_y\hat{\mathbf{k}})\, dA + \text{etc.},\end{aligned}$$

apply the divergence theorem to each integral, and so on.

(c) Deduce, from equation (9.1), that $\int_S \hat{\mathbf{n}}\, dA = \mathbf{0}$ for every S. Verify this result, by direct integration, for the case where S is the surface of the rectangular prism $0 \le x \le a$, $0 \le y \le b$, $0 \le z \le c$. NOTE: Just as the global statement $\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\, dA = \int_V \nabla\cdot\mathbf{v}\, dV$ corresponds to the local statement

$$\nabla\cdot\mathbf{v} = \lim_{V\to 0}\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\, dA}{V}$$

(V = the volume enclosed by S), (9.1) and (9.2)correspond to the local statements

$$\nabla u = \lim_{V\to 0}\frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V} \tag{9.3}$$

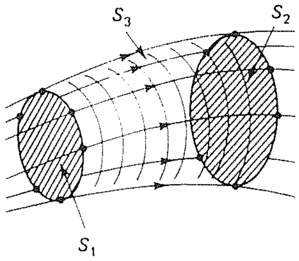
and

$$\nabla\times\mathbf{v} = \lim_{V\to 0}\frac{\int_S \hat{\mathbf{n}}\times\mathbf{v}\, dA,}{V} \tag{9.4}$$

respectively.

10. (*Continuity equation*) Let S_1 be a completely permeable plane surface within a steady fluid flow field. Let the velocity field be $\mathbf{v}(x, y, z)$, and let the mass density field be $\sigma(x, y, z)$. Then the set of streamlines that pass through the set points on the edge of S_1form a "streamtube," as sketched in the accompanying figure. Let S_2 be another plane cross section of the streamtube an arbitrary distance downstream of S_1, and let the areas of S_1, S_2 be A_1, A_2, respectively. Recalling the continuity equation (29), namely $\nabla\cdot(\sigma\mathbf{v}) = 0$ (since the field is steady, so that $\partial\sigma/\partial t = 0$), integrate this equation over a control volume V which is bounded by S_1, S_2, and the streamtube surface S_3 between S_1 and S_2, apply the divergence theorem, and thus show that

$$\int_{S_1} \hat{\mathbf{n}}\cdot(\sigma\mathbf{v})\, dA + \int_{S_2} \hat{\mathbf{n}}\cdot(\sigma\mathbf{v})\, dA = 0. \tag{10.1}$$



If σ and $\mathbf{v}$ are constant over S_1 and S_2, and S_1, S_2 are normal to the flow, show that (10.1) reduces to the simpler form

$$\sigma_1 A_1 V_1 = \sigma_2 A_2 V_2, \tag{10.2}$$

often given in fluid mechanics texts, where $\|\mathbf{v}\| = V_1$ on S_1 and $\|\mathbf{v}\| = V_2$ on S_2.

11. (*Archimedes's principle*) Consider a solid body immersed in a fluid of uniform mass density σ. Archimedes's principle states that the net pressure force on the body is upward (i.e., it is a buoyant force) and equal to the weight of the fluid displaced by the body. Derive Archimedes's principle using the methods of this section. HINT: If we measure z upwards from the surface of the fluid, then the static pressure distribution in the fluid is $p(x, y, z) = -\sigma g z$. Write the net force $\mathbf{F}$ on the submerged body as an integral over its surface S, and then transform that to a volume integral using (9.1) in Exercise 9.

12. Verify the divergence theorem.

(a) $\mathbf{v} = 3r^2\hat{\mathbf{e}}_r - r\hat{\mathbf{e}}_\theta + 2\hat{\mathbf{e}}_z$, V: the cylinder $r \le 4$, $0 \le \theta < 2\pi$, $0 \le z \le 5$

(b) $\mathbf{v} = z\hat{\mathbf{e}}_z$, V: the cylinder $r \le 2$, $0 \le \theta < 2\pi$, $-3 \le z \le 6$

(c) $\mathbf{v} = rz^2\hat{\mathbf{e}}_\theta$, V: the half-cylinder $r \le 4$, $\pi/2 \le \theta \le 3\pi/2$,

$0 \le z \le 3$
(d) $\mathbf{v} = z^2\hat{\mathbf{e}}_r + 6\hat{\mathbf{e}}_z$, $\mathcal{V}$: the hollow cylinder $2 \le r \le 3$, $0 \le \theta < 2\pi, 1 \le z \le 5$
(e) $\mathbf{v} = z^2\hat{\mathbf{e}}_z$, $\mathcal{V}$: the cone $r \le 2z, 0 \le \theta < 2\pi, 0 \le z \le 3$
(f) $\mathbf{v} = \rho^2\hat{\mathbf{e}}_\rho$, $\mathcal{V}$: the sphere $\rho \le a$
(g) $\mathbf{v} = \rho^3\hat{\mathbf{e}}_\rho$, $\mathcal{V}$: the hemisphere $\rho \le a$, $0 \le \phi < \pi/2$, $0 \le \theta < 2\pi$
(h) $\mathbf{v} = \rho^2 \sin\phi \sin\theta \hat{\mathbf{e}}_\theta$, $\mathcal{V}$: the region $2 \le \rho \le 3$, $0 \le \phi < \pi/2, 0 \le \theta \le \pi/2$
(i) $\mathbf{v} = \dfrac{1}{\rho}\hat{\mathbf{e}}_\rho + \rho^3\hat{\mathbf{e}}_\phi - \rho^2 \sin^2\phi\hat{\mathbf{e}}_\theta$, $\mathcal{V}$: the hollow sphere $2 \le \rho \le 5$

16.9 Stokes's Therem

Following Newton's death in 1727 the mathematical scene in the British Isles was relatively quiet for almost 100 years. There then appeared a number of exceptional mathematical physicists, beginning with *George Green* (1793–1841) and *William R. Hamilton* (1805–1865), and followed by the "Cambridge school," which included *Sir G. Gabriel Stokes* (1819–1903), *Lord Kelvin* (Sir William Thomson; 1824–1907), *James Clerk Maxwell* (1831–1879), and *Lord Rayleigh* (John William Strutt; 1842–1919).

In this section we examine two important and closely related theorems, one due to Stokes and one due to Green. Both of these theorems involve line integrals so we begin by reviewing and extending our study of line integrals, which began in Section 15.2.

16.9.1. Line integrals. We met the line integral $\int_C f\,ds$ and showed how to evaluate it by parametrizing the curve C according to $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$ for $a \le \tau \le b$. Specifically,

$$\int_C f\,ds = \int_a^b f(x(\tau), y(\tau), z(\tau))\sqrt{\mathbf{R}'(\tau) \cdot \mathbf{R}'(\tau)}\,d\tau, \tag{1}$$

where $\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}$. That is, by parametrizing we reduce the line integral $\int_C f\,ds$ on a curve C to an "ordinary" integral of the form $\int_a^b F(\tau)\,d\tau$ on a segment of a τ axis. Actually, in engineering science applications it is more common to meet line integrals in a different form, namely, in the form

$$\int_C \mathbf{v} \cdot d\mathbf{R}, \tag{2}$$

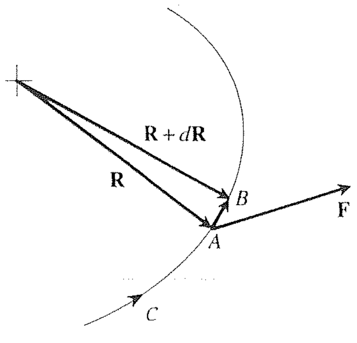


Figure 1. Work.

where $\mathbf{v}$ is a given vector field defined in some region containing C, and $\mathbf{R}$ is the position vector from some reference point, or origin, to points on the curve C. For example, if a particle is subjected to a force $\mathbf{F}$ as it moves along a curve C, then the work dW done by $\mathbf{F}$, as the particle moves an infinitesimal distance along C, say from A to B (Fig. 1) is

$$dW = \mathbf{F} \cdot \mathbf{AB} = \mathbf{F} \cdot d\mathbf{R}. \tag{3}$$

Evidently, then, the net work done in traversing the entire curve C is

$$W = \int_C \mathbf{F} \cdot d\mathbf{R}. \tag{4}$$

For instance, $\mathbf{F}(x, y, z)$ might be a gravitational force field induced by point or distributed systems of mass. Since the curve C, in Fig. 1, is actually traversed by the particle, we sometimes call it a *path* instead of a curve.

That the two forms $\int_C f\, ds$ and $\int_C \mathbf{v} \cdot d\mathbf{R}$ are equivalent can be seen by expressing

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_C \mathbf{v} \cdot \frac{d\mathbf{R}}{ds}\, ds = \int_C f\, ds, \tag{5}$$

where f is $\mathbf{v} \cdot d\mathbf{R}/ds$.* However, there is one difference to keep in mind. In $\int_C \mathbf{v} \cdot d\mathbf{R}$ the curve C is said to be **oriented**; that is, we have a specific direction of traversal in mind, and we denote it graphically by an arrowhead, as we did in Fig. 1. In that illustration, observe that the work done in moving from A to B, $dW = \mathbf{F} \cdot \mathbf{AB}$, is the negative of the work done if instead we moved from B to A, $dW = \mathbf{F} \cdot \mathbf{BA}$, because $\mathbf{BA} = -\mathbf{AB}$. Thus, if we denote an oriented curve as C, and the same curve but oppositely oriented as "$-C$," then

$$\int_{-C} \mathbf{v} \cdot d\mathbf{R} = -\int_C \mathbf{v} \cdot d\mathbf{R}. \tag{6}$$

In contrast, the curve C in the form $\int_C f\, ds$ is *not* oriented because ds is arc length, which is always positive. That claim is consistent with (1) because we see in (1) that $ds = \sqrt{\mathbf{R}'(\tau) \cdot \mathbf{R}'(\tau)}\, d\tau$, where the positive square root is understood, and where each $d\tau$ is positive because $b \geq a$. Thus, in $\int_C f\, ds$ the curve C is not oriented, and in $\int_C \mathbf{v} \cdot d\mathbf{R}$ it is.

Just as we evaluate $\int_C f\, ds$ by parametrizing C as in (1), we evaluate $\int_C \mathbf{v} \cdot d\mathbf{R}$ in the same manner. We begin by parametrizing C by

$$\left.\begin{aligned} x &= x(\tau) \\ y &= y(\tau) \\ z &= z(\tau) \end{aligned}\right\} \tau : a \rightarrow b, \tag{7}$$

where we write $\tau : a \rightarrow b$ rather than $a \leq \tau \leq b$ because the value a (at the initial point on C) can be less than *or* greater than the value b (at the terminal point on C) because of the orientation of C. Then,

$$\begin{aligned} \int_C \mathbf{v} \cdot d\mathbf{R} &= \int_C (v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}} + v_z \hat{\mathbf{k}}) \cdot (dx \hat{\mathbf{i}} + dy \hat{\mathbf{j}} + dz \hat{\mathbf{k}}) \\ &= \int_C v_x(x, y, z)\, dx + v_y(x, y, z)\, dy + v_z(x, y, z)\, dz \end{aligned}$$

*We can regard the position vector $\mathbf{R}$ as a function of the arc length s along C by parametrizing C by $x = x(s)$, $y = y(s)$, $z = z(s)$ since then $\mathbf{R}(x(s), y(s), z(s))$ is a function of s.

$$= \int_a^b \left[v_x(x(\tau), y(\tau), z(\tau)) \frac{dx}{d\tau} + v_y(x(\tau), y(\tau), z(\tau)) \frac{dy}{d\tau} \right.$$
$$\left. + v_z(x(\tau), y(\tau), z(\tau)) \frac{dz}{d\tau} \right] d\tau. \tag{8}$$

EXAMPLE 1. Evaluate $\int_C \mathbf{v} \cdot d\mathbf{R}$, where $\mathbf{v} = xz^2\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + 2y\hat{\mathbf{k}}$, and where C is the bent line from $A = (2,1,3)$ to $B = (1,2,1)$ to $D = (3,4,5)$, shown schematically in Fig. 2. First, break C into the two parts C_1 and C_2, and write

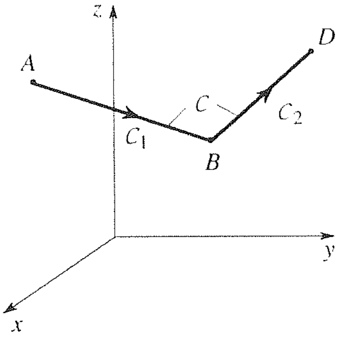


Figure 2. The curve C.

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_{C_1} \mathbf{v} \cdot d\mathbf{R} + \int_{C_2} \mathbf{v} \cdot d\mathbf{R},$$
$$= \int_{C_1} xz^2\, dx - 3\, dy + 2y\, dz + \int_{C_2} xz^2\, dx - 3\, dy + 2y\, dz, \tag{9}$$

which step is permissible for $\int_C \mathbf{v} \cdot d\mathbf{R}$ just as it is for $\int_C f\, ds$ [recall equations (12) in Section 15.2]. Next, use (7) in Section 15.2 to parametrize C_1 and C_2:

$$C_1: \quad \begin{aligned} x &= 2 + (1-2)\tau = 2 - \tau, & x'(\tau) &= -1 \\ y &= 1 + (2-1)\tau = 1 + \tau, & y'(\tau) &= 1 \\ z &= 3 + (1-3)\tau = 3 - 2\tau, & z'(\tau) &= -2 \end{aligned} \tag{10a}$$

and

$$C_2: \quad \begin{aligned} x &= 1 + (3-1)\tau = 1 + 2\tau, & x'(\tau) &= 2 \\ y &= 2 + (4-2)\tau = 2 + 2\tau, & y'(\tau) &= 2 \\ z &= 1 + (5-1)\tau = 1 + 4\tau, & z'(\tau) &= 4 \end{aligned} \tag{10b}$$

as τ goes from 0 to 1 (i.e., $\tau : 0 \to 1$). Finally, expressing dx, dy, dz in (9) as $x'(\tau)\, d\tau$, $y'(\tau)\, d\tau$, $z'(\tau)\, d\tau$, respectively, and using (10), we have

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_0^1 [(2-\tau)(3-2\tau)^2(-1) - 3(1) + 2(1+\tau)(-2)]\, d\tau$$
$$+ \int_0^1 [(1+2\tau)(1+4\tau)^2(2) - 3(2) + 2(2+2\tau)(4)]\, d\tau$$
$$= -\frac{97}{6} + \frac{202}{3} = \frac{307}{6}. \quad \blacksquare \tag{11}$$

Recall from Section 15.2 that a curve is said to be *closed* if its endpoints coincide. If C is closed, then we usually write

$$\oint_C \mathbf{v} \cdot d\mathbf{R} \tag{12}$$

in place of $\int_C \mathbf{v} \cdot d\mathbf{R}$. One might think that we will always obtain $\oint_C \mathbf{v} \cdot d\mathbf{R} = 0$ because the endpoints of C coincide, just as it is true that $\int_a^a f(x)\, dx = 0$ for an integration on the x axis, but that is not necessarily the case. For instance, suppose we close the contour C in Example 1 by adding a straight line segment from D to A (Fig. 2). Call that segment C_3. We find (Exercise 1) that $\int_{C_3} \mathbf{v} \cdot d\mathbf{R} = -\frac{259}{6}$ so

$$\oint_C \mathbf{v} \cdot d\mathbf{R} = \int_{ABDA} \mathbf{v} \cdot d\mathbf{R} = -\frac{97}{6} + \frac{202}{3} - \frac{259}{6} = 8, \tag{13}$$

which is not zero. Of course, a line integral around a closed path *may* be zero, but it need not.

Two examples of line integrals around closed contours are as follows. First, *Ampère's law*,

$$\oint_C \mathbf{H} \cdot d\mathbf{R} = I, \tag{14}$$

states that the line integral of the magnetic field intensity $\mathbf{H}$ (amperes/meter) around a closed curve C equals the current I (amperes) flowing through C. That is, think of C as made of a stiff wire, dipped in soap water so that there is a soap film S with C as its perimeter. Then I is the current crossing the surface S. Second, in fluid mechanics the line integral

$$\oint_C \mathbf{v} \cdot d\mathbf{R} = \Gamma \tag{15}$$

defines the *circulation* Γ (meters^2/second) of the fluid velocity field $\mathbf{v}$ around the contour C; Γ is important in aerodynamics because of its prominent role in the *Kutta–Joukowski formula*,

$$L = \sigma U \Gamma, \tag{16}$$

for the lift L per unit span of an airfoil (i.e., a wing) due to a fluid flow of velocity U and density σ. In (16), Γ is the circulation around any closed clockwise contour that encloses the airfoil such as the ellipse shown in Fig. 3. That is, consider a simple airfoil consisting of a flat plate at incidence α, normal to the paper, a cross section of which is shown in Fig. 3. Rather than the plate moving leftward with flight speed U, it is equivalent to consider the plate as stationary in a rightward flow of speed U, as in a wind tunnel. If $\alpha = 0$ the streamlines are simply horizontal straight lines, but as α is increased the streamlines distort more and more (somewhat as the semicircular bump causes a distortion of the otherwise-horizontal streamlines in Fig. 4 of Section 16.2). Although the streamline pattern is not shown here, in Fig. 3, it could be found by the method of conformal mapping.* In any case, the result of such calculation is that the velocities above the plate are larger than those below the plate, as sketched at points A and B in the figure. Thus, for $d\mathbf{R}$'s of equal length, the $\mathbf{v} \cdot d\mathbf{R}$ contribution to Γ from point A is positive and larger than the negative contribution from point B so that a positive circulation Γ is established around C by virtue of the inclination α of the plate. According to (16), that circulation gives rise to a lift force. Even if you haven't studied fluid mechanics or aerodynamics you may have studied the Bernoulli equation (relating pressure and velocity) in a

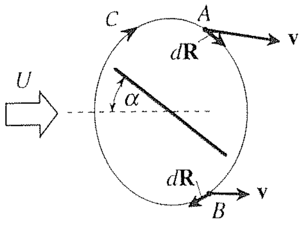


Figure 3. Lift on an airfoil.

*In fact, that solution is outlined, for the case where $\alpha = 90°$, in Exercise 6 in Section 23.6, and it could be modified for $\alpha \neq 90°$.

physics course. Since Bernoulli's equation states that high velocity gives low pressure and low velocity gives high pressure, it follows – from the fact that the velocity is greater above the plate than below it – that the pressure is lower above the plate than below it, and that pressure difference establishes the lift L.

16.9.2. Stokes's theorem. Recall from the Closure in Section 16.8 that the fundamental theorem of the integral calculus,

$$\int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b}, \tag{17}$$

gives the integral along a (one-dimensional) line segment as an evaluation at the (zero-dimensional) boundary of that line, namely, its endpoints, and that the Gauss divergence theorem gives the integral over a (three-dimensional) region as an integration over the (two-dimensional) boundary of that region, namely, its surface. Filling the gap between those two results, Stokes's theorem gives the integral over a two-dimensional open surface $\mathcal{S}$ in terms of a line integral around its one-dimensional edge curve C. Here we distinguish a **closed surface**, which encloses a volume (as does the surface $\mathcal{S}$ in the divergence theorem), from an **open surface**, which does not. For instance, the surface of a basketball is closed and the surface defined by a potato chip is open.

THEOREM 16.9.1 *Stokes's Theorem*
Let $\mathbf{v}$ be a C^1 vector field defined in a region $\mathcal{R}$ in 3-space. Let $\mathcal{S}$ be an open piecewise smooth orientable surface within $\mathcal{R}$, and let the edge of $\mathcal{S}$ be a piecewise smooth simple closed curve C. Then

$$\boxed{\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA = \oint_C \mathbf{v}\cdot d\mathbf{R},} \tag{18}$$

where $\hat{\mathbf{n}}$ is a unit normal to $\mathcal{S}$, the orientation of C and the direction of $\hat{\mathbf{n}}$ being in accordance with the right-hand rule (Fig. 4).

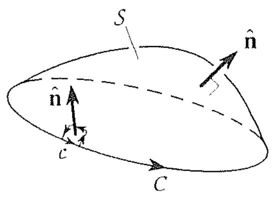


Figure 4. $\mathcal{S}, C$, and $\hat{\mathbf{n}}$ in (18).

Before discussing the proof of Stokes's theorem we need to explain our statement in the theorem about the right-hand rule. First, observe that we refrain from calling $\hat{\mathbf{n}}$ the "outward" or "inward" normal to $\mathcal{S}$. If $\mathcal{S}$ were a closed surface, such as the surface of a sphere or cube, it would make sense to speak of an outward (or inward) normal to $\mathcal{S}$ since $\mathcal{S}$ would "inherit" an orientation from the volume $\mathcal{V}$ that it enclosed: normal vectors on $\mathcal{S}$ directed into $\mathcal{V}$ would be called inward, and those directed out of $\mathcal{V}$ would be called outward. But in Stokes's theorem $\mathcal{S}$ is not a closed surface so the terms "outward" and "inward" do not apply.

The idea, then, is that $\hat{\mathbf{n}}$, in (18), can be either of the two (oppositely directed) fields of normals on $\mathcal{S}$. Similarly, C can be oriented in either of two possible directions. Selecting one of those two orientations for C, arbitrarily, we then choose between the two possible normals so that $\hat{\mathbf{n}}$ and C are in accordance with the right-hand rule. That is, if we consider a small contour C', part of which coincides with (and is oriented in the same way as) the boundary curve C, then applying the right-hand rule to C' yields a normal $\hat{\mathbf{n}}$ (Fig. 4). We then continue this normal over the rest of S. For example, if the surface S is a flat surface in the x, y plane, then we can have $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$ and C counterclockwise, *or* $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$ and C clockwise.

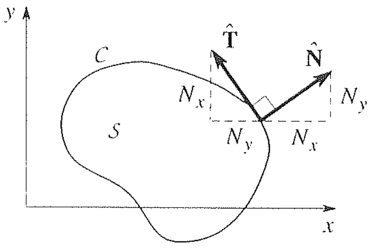


Figure 5. Plane region.

Proof: Let us limit our proof of (18) to the case where S is flat, in which case S can be assumed to lie in the plane of the paper as in Fig. 5; extension from the plane case to the general case will be left for the exercises.

Recall the two-dimensional divergence theorem given by (47) in Section 16.8.2. In the present case the region "$\mathcal{R}$" in (47) is S, and in place of "v" and "$\hat{\mathbf{n}}$" it will be convenient to use the letters $\mathbf{V}$ and $\hat{\mathbf{N}}$. Thus,

$$\int_S \nabla \cdot \mathbf{V}\, dA = \int_C \hat{\mathbf{N}} \cdot \mathbf{V}\, ds. \tag{19}$$

Writing out the integrands gives

$$\int_S \left(\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} \right) dA = \int_C (N_x V_x + N_y V_y)\, ds, \tag{20}$$

where $\mathbf{V} = V_x \hat{\mathbf{i}} + V_y \hat{\mathbf{j}}$ and $\hat{\mathbf{N}} = N_x \hat{\mathbf{i}} + N_y \hat{\mathbf{j}}$ (Fig. 5). If we define a new vector $\mathbf{v} = v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}} \equiv -V_y \hat{\mathbf{i}} + V_x \hat{\mathbf{j}}$, then (20) becomes

$$\begin{aligned}\int_S \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) dA &= \int_C (N_x v_y - N_y v_x)\, ds \\ &= \int_C (v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}}) \cdot (-N_y \hat{\mathbf{i}} + N_x \hat{\mathbf{j}})\, ds \\ &= \int_C \mathbf{v} \cdot \hat{\mathbf{T}}\, ds, \end{aligned} \tag{21}$$

where $\hat{\mathbf{T}} = -N_y \hat{\mathbf{i}} + N_x \hat{\mathbf{j}}$ is the unit counterclockwise tangent vector to C. Observe that the integrand on the left is the z component of $\nabla \times \mathbf{v}$, namely, $\hat{\mathbf{k}} \cdot \nabla \times \mathbf{v}$, and that $\hat{\mathbf{T}}\, ds = d\mathbf{R}$ along C. Finally, $\hat{\mathbf{k}}$ is identical to the $\hat{\mathbf{n}}$ in Stokes's theorem so (21) becomes

$$\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_C \mathbf{v} \cdot d\mathbf{R},$$

which was to be proved.

Here S was flat and $\mathbf{v}$ was only a two-dimensional field. A proof for the general case, where S need not be flat and $\mathbf{v}$ is three-dimensional, is outlined in the exercises. ■

How can we understand Stokes's theorem from a physical point of view? For definiteness, think of $\mathbf{v}$ as a fluid velocity field (which we can do even if it is not).

Recall, from our discussion of fluid mechanics and the curl, that $\boldsymbol{\nabla} \times \mathbf{v}$ is twice the fluid particle angular velocity and is known as the vorticity and denoted as $\boldsymbol{\Omega}$. Thus, $\int_S \hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \times \mathbf{v}\, dA = \int_S \hat{\mathbf{n}} \cdot \boldsymbol{\Omega}\, dA$ is the vorticity flux across S. Furthermore, $\oint_C \mathbf{v} \cdot d\mathbf{R}$ is the circulation around C. Thus, in fluid mechanics terminology, Stokes's theorem (18) states that *the circulation* Γ *of the flow field* $\mathbf{v}$, *around* C, *equals the vorticity flux through* S.*

EXAMPLE 2. *Vorticity Flux.* Consider the plane flow field $\mathbf{v} = \kappa y \hat{\mathbf{i}}$, where κ is some positive constant. The flow, called a *shear flow*, is shown in Fig. 6. Let S be a generic region in the x, y plane, and let the edge curve be clockwise, say, so that $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$. Then $\boldsymbol{\Omega} = \boldsymbol{\nabla} \times \mathbf{v} = -\kappa \hat{\mathbf{k}}$ is itself a "flow," a uniform flow, of magnitude κ, directed into the paper. Thus, the vorticity flux through S is simply κ times the area A of S, and that is exactly what the left-hand side of (18) gives,

$$\int_S \hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \times \mathbf{v}\, dA = \int_S \hat{\mathbf{n}} \cdot \boldsymbol{\Omega}\, dA = \int_S -\hat{\mathbf{k}} \cdot (-\kappa \hat{\mathbf{k}})\, dA = \kappa A. \quad \blacksquare$$

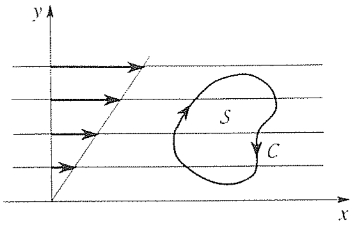


Figure 6. Shear flow.

EXAMPLE 3. *Verification of Stokes's Theorem.* Verify Stokes's theorem for the case where $\mathbf{v} = xz\hat{\mathbf{j}}$, and where S is the surface $z = 4 - y^2$, cut off by the planes $x = 0$, $z = 0$, and $y = x$, with C oriented as shown in Fig. 7; we have broken C into the three parts, C_1, C_2, C_3. Let us work out the left-hand side of (18) first. Using the gradient method to find $\hat{\mathbf{n}}$ [equation (20) in Section 16.8], and the formula

$$dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy \tag{22}$$

for dA [equation (11) in Section 16.8, where S is the surface $z = f(x, y) = 4 - y^2$ in this case], we have

$$\boldsymbol{\nabla} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ 0 & xz & 0 \end{vmatrix} = -x\hat{\mathbf{i}} + z\hat{\mathbf{k}},$$

$$\hat{\mathbf{n}} = \frac{\boldsymbol{\nabla}(z + y^2)}{\|\boldsymbol{\nabla}(z + y^2)\|} = \frac{2y\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{4y^2 + 1}},$$

$$\hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \times \mathbf{v} = \frac{z}{\sqrt{4y^2 + 1}},$$

$$dA = \sqrt{1 + (0)^2 + (-2y)^2}\, dx\, dy = \sqrt{1 + 4y^2}\, dx\, dy \tag{23}$$

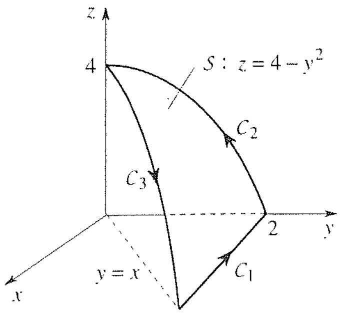


Figure 7. Example 3.

so

$$\int_S \hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \times \mathbf{v}\, dA = \int_0^2 \int_0^y \left. \frac{z}{\sqrt{4y^2 + 1}} \right|_{z = 4 - y^2} \sqrt{1 + 4y^2}\, dx\, dy$$

$$= \int_0^2 \int_0^y (4 - y^2)\, dx\, dy = 4, \tag{24}$$

*Actually, (18) is not "Stokes's theorem" it is only the formula within Stokes's theorem, but for brevity it is convenient to refer to it as Stokes's theorem.

where the x, y integration limits correspond to the triangular "shadow" of $\mathcal{S}$ down onto the x, y plane. Of course, the integration is actually on $\mathcal{S}$ nonethless; it has simply been "referred" to the x, y plane by virtue of the relation (22) between dA on $\mathcal{S}$ and $dx\,dy$ in the x, y plane.

Turning to the right-hand side of (18), we have

$$\begin{aligned}\oint_C \mathbf{v}\cdot d\mathbf{R} &= \oint_C xz\hat{\mathbf{j}}\cdot(dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}) = \oint_C xz\,dy \\ &= \oint_{C_1} xz\,dy + \oint_{C_2} xz\,dy + \oint_{C_3} xz\,dy. \end{aligned} \tag{25}$$

The first integral is zero because $z = 0$ on C_1, and the second is zero because $x = 0$ on C_2. To evaluate the C_3 integral, we can parametrize C_3 as follows. Let $x = \tau$, say. Then $y = \tau$ because $y = x$ on C_3, and $z = 4 - \tau^2$ because $z = 4 - y^2$ on C_3. Finally, τ goes from 0 to 2 because $x = \tau$ and x goes from 0 to 2 as we move along C_3 from the initial point $(0, 0, 4)$ to the final point $(2, 2, 0)$. Thus, (25) gives

$$\oint_C \mathbf{v}\cdot d\mathbf{R} = 0 + 0 + \int_0^2 \tau(4 - \tau^2)\,d\tau = 4, \tag{26}$$

in agreement with (24). ▮

EXAMPLE 4. *Ampère's Law.* As a physical application of Stokes's theorem, let us derive one of the celebrated Maxwell's equations of electromagnetic field theory. We consider a region of 3-space in which there is a magnetic field with magnetic field intensity $\mathbf{H}$ (amperes/meter), a charge density field σ (coulombs/meter^3), and a velocity field $\mathbf{v}$ (meters/second). In general, the scalar field σ and the vector fields $\mathbf{H}$ and $\mathbf{v}$ will vary with x, y, z, and the time t, but any dependence on t will not be relevant here.

We begin with Ampère's law,

$$\oint_C \mathbf{H}\cdot d\mathbf{R} = I, \tag{27}$$

which states that the line integral of the magnetic field intensity $\mathbf{H}$ around any closed curve C equals the current I (amperes) passing through any control surface $\mathcal{S}$ (i.e., a mathematical surface, not a physical surface) having C as its boundary. (Note that when we say "any" surface or "any" curve C it is understood that they be suitably decent, namely, that they satisfy the conditions on $\mathcal{S}$ and C in Stokes's theorem.)

Since I is the flux of charge across $\mathcal{S}$ we can express it as

$$I = \int_S \sigma\,\hat{\mathbf{n}}\cdot\mathbf{v}\,dA = \int_S \hat{\mathbf{n}}\cdot(\sigma\mathbf{v})\,dA, \tag{28}$$

where $\hat{\mathbf{n}}$ is the unit normal to $\mathcal{S}$ in whichever of the two directions is chosen as the direction in which I is positive. As a partial check on (28), note that its units, (coulombs/m^3)(m/sec)(m^2), do indeed give coulombs/sec.

Using Stokes's theorem to express

$$\oint_C \mathbf{H}\cdot d\mathbf{R} = \int_S \hat{\mathbf{n}}\cdot\nabla\times\mathbf{H}\,dA, \tag{29}$$

(27)–(29) give

$$\int_S \hat{\mathbf{n}}\cdot(\nabla\times\mathbf{H}-\mathbf{J})\,dA = 0, \tag{30}$$

where it is traditional to combine the product $\sigma\mathbf{v}$ as a single field $\mathbf{J} = \sigma\mathbf{v}$, known as the current density. Finally, since (30) holds for *any* surface $\mathcal{S}$, it follows that $\nabla\times\mathbf{H}-\mathbf{J}$ must be identically zero in the region so

$$\nabla\times\mathbf{H} = \mathbf{J}, \tag{31}$$

which is one of Maxwell's equations.

Why does it follow from (30) that $\nabla\times\mathbf{H}-\mathbf{J} = 0$? Can't the integral be zero by virtue of cancellation, without the integrand being *identically* zero? For a specific $\mathcal{S}$, yes, but not for all possible choices of $\mathcal{S}$. For suppose that $\nabla\times\mathbf{H}-\mathbf{J}$ is nonzero at some point P; let it be nonzero vector $\mathbf{a}$ at P. Then let us choose $\mathcal{S}$ to be a circular disk with its center at P, and let its normal $\hat{\mathbf{n}}$ be aligned with $\mathbf{a}$ (Fig. 8). Of course $\nabla\times\mathbf{H}-\mathbf{J} = \mathbf{a}$ only at P; it varies on $\mathcal{S}$. But if it is continuous, which we assume, then it must be possible to choose the radius ϵ of $\mathcal{S}$ small enough so that $(\nabla\times\mathbf{H}-\mathbf{J})\cdot\hat{\mathbf{n}} > 0$ everywhere on $\mathcal{S}$, in which case the integral in (30) is positive rather than zero. This contradiction of (30) shows that it must not be possible for $\nabla\times\mathbf{H}-\mathbf{J}$ to be nonzero anywhere in the field.

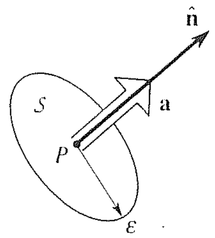


Figure 8. Choosing $\mathcal{S}$.

COMMENT 1. The logic used to infer (31) from (30) is very similar to that used to infer (28) from (27) and also (38) from (37) in Section 16.8. We suggest that you review those two examples in Section 16.8

COMMENT 2. Understand that (31) is a *field equation*, specifically a partial differential equation relating the vector fields $\mathbf{H}$ and $\mathbf{J}$ at each point in the field. It happens to be a vector equation or, equivalently, the three scalar PDE's

$$\begin{aligned}
\frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= J_x,\\
\frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} &= J_y,\\
\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= J_z,
\end{aligned} \tag{32}$$

whereas the field equations (28) and (38) in Section 16.8 were single scalar PDE's. ∎

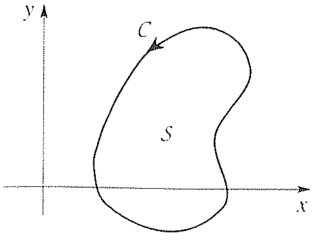


Figure 9. Stokes's theorem for plane case.

16.9.3. Green's theorem. Let us apply Stokes's theorem to the plane case where $\mathbf{v} = P(x,y)\hat{\mathbf{i}} + Q(x,y)\hat{\mathbf{j}}$, and where $\mathcal{S}$ is a flat surface in the x,y plane, and let C be counterclockwise (Fig. 9). Then $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$ and $\nabla\times\mathbf{v} = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\hat{\mathbf{k}}$ so we have the following result, which is known as **Green's theorem**.

THEOREM 16.9.2 *Green's Theorem*
Let a C^1 vector field $\mathbf{v} = P(x,y)\hat{\mathbf{i}} + Q(x,y)\hat{\mathbf{j}}$ be defined in a region $\mathcal{R}$ in 2-space. Let $\mathcal{S}$ be a region within $\mathcal{R}$, and let the edge of $\mathcal{S}$ be a piecewise smooth simple closed curve C, oriented counterclockwise. Then

$$\boxed{\int_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \oint_C P\,dx + Q\,dy.} \tag{33}$$

Since Green's theorem is really but a special case of Stokes's theorem, let us limit our discussion of it to an illustrative example, with additional material reserved for the exercises.

EXAMPLE 5. Verify (33) for the case where $P(x,y) = xy^3$, and $Q(x,y) = x^2 - y^2$, and where $\mathcal{S}$ is as shown in Fig. 10. Then

$$\begin{aligned}
\int_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA &= \int_0^2 \int_{y/2}^1 (2x - 3xy^2)\,dx\,dy \\
&= \int_0^2 \left(x^2 - \frac{3}{2}x^2y^2 \right)\Big|_{x=y/2}^{x=1} dy \\
&= \int_0^2 \left(1 - \frac{3}{2}y^2 - \frac{y^2}{4} + \frac{3}{2}\frac{y^2}{4}y^2 \right) dy \\
&= \left(y - \frac{1}{2}y^3 - \frac{1}{12}y^3 + \frac{3}{40}y^5 \right)\Big|_0^2 = -\frac{4}{15}.
\end{aligned} \tag{34}$$

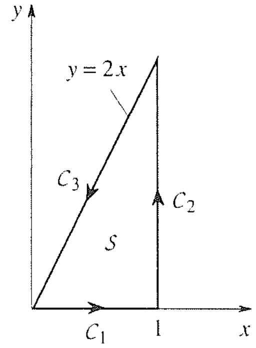


Figure 10. The region $\mathcal{S}$.

Breaking C into $C_1 + C_2 + C_3$, we have

$$\begin{aligned}
\oint_C P\,dx + Q\,dy &= \int_{C_1} xy^3\,dx + (x^2 - y^2)\,dy + \int_{C_2} xy^3\,dx + (x^2 - y^2)\,dy \\
&\quad + \int_{C_3} xy^3\,dx + (x^2 - y^2)\,dy \\
&= \int_{C_2} (x^2 - y^2)\,dy + \int_{C_3} xy^3\,dx + (x^2 - y^2)\,dy
\end{aligned} \tag{35}$$

since $y = 0$ on C_1, $dy = 0$ on C_1, and $dx = 0$ on C_2. Finally, since $x = 1$ on C_2, and $y = 2x$ on C_3, (35) gives

$$\oint_C P\,dx + Q\,dy = \int_0^2 (1 - y^2)\,dy + \int_1^0 [8x^4 + (x^2 - 4x^2)(2)]\,dx = -\frac{4}{15}, \tag{36}$$

in agreement with (34).

COMMENT. Why didn't we parametrize C_2 and C_3 to evaluate the final integrals in (35)? Actually, we did; in C_2 we used y as the parameter for C_2 (i.e., we used $y = \tau$ but didn't bother changing the name from y to τ), and we used x as the parameter for C_3. If you prefer, use τ explicitly. ▮

16.9.4. Non-Cartesian coordinates. (Optional) Let us illustrate the occurrence of non-Cartesian coordinates, in Stokes's theorem, with two examples.

EXAMPLE 6. *Cylindrical Coordinates.* Verify Stokes's theorem for the case where

$$\mathbf{v} = -r^2\cos\theta\hat{\mathbf{e}}_r + r^2\hat{\mathbf{e}}_\theta + r\sin\theta\hat{\mathbf{e}}_z, \tag{37}$$

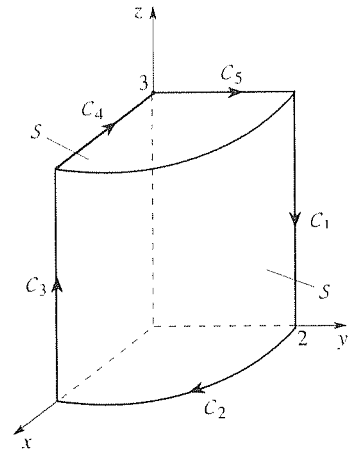


Figure 11. S and C for Example 6.

where S is as shown in Fig. 11 (essentially, a quarter of a soup can but with no bottom), and where C is oriented as shown in the figure.

Let us do the surface integral first. From the cylindrical coordinate expression for $\nabla\times\mathbf{v}$ in Section 16.7,

$$\nabla\times\mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta + (3r - r\sin\theta)\hat{\mathbf{e}}_z. \tag{38}$$

On the curved part of the surface $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_r$ (in accordance with the orientation of C) and $dA = r\,d\theta\,dz$ (for a constant-r surface), and on the flat top $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_z$ and $dA = r\,dr\,d\theta$ (for a constant-z surface) so

$$\begin{aligned}\int_S \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA &= \int_0^3\int_0^{\pi/2}(-\hat{\mathbf{e}}_r)\cdot\nabla\times\mathbf{v}\,r\,d\theta\,dz\Big|_{r=2}\\ &\quad+\int_0^{\pi/2}\int_0^2(-\hat{\mathbf{e}}_z)\cdot\nabla\times\mathbf{v}\;r\,dr\,d\theta\Big|_{z=3}\\ &= -2\int_0^3\int_0^{\pi/2}\cos\theta\,d\theta\,dz - \int_0^{\pi/2}\int_0^2(3r^2 - r^2\sin\theta)\,dr\,d\theta\\ &= -\frac{10}{3} - 4\pi.\end{aligned} \tag{39}$$

Turning to the line integral, we obtain

$$\begin{aligned}\oint_C \mathbf{v}\cdot d\mathbf{R} &= \oint_C(-r^2\cos\theta\hat{\mathbf{e}}_r + r^2\hat{\mathbf{e}}_\theta + r\sin\theta\hat{\mathbf{e}}_z)\cdot(dr\,\hat{\mathbf{e}}_r + r\,d\theta\,\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z)\\ &= \oint_C -r^2\cos\theta\,dr + r^3\,d\theta + r\sin\theta\,dz\\ &= \int_3^0 r\sin\theta\,dz\Big|_{r=2,\theta=\pi/2} + \int_{\pi/2}^0 r^3\,d\theta\Big|_{r=2,z=0} + \int_0^3 r\sin\theta\,dz\Big|_{r=2,\theta=0}\\ &\quad+\int_2^0(-r^2\cos\theta)\,dr\Big|_{\theta=0,z=3} + \int_0^2(-r^2\cos\theta)\,dr\Big|_{\theta=\pi/2,z=3}\\ &= -6 - 4\pi + 0 + \frac{8}{3} + 0 = -\frac{10}{3} - 4\pi,\end{aligned} \tag{40}$$

in agreement with (39). Following the third equality in (40), the five integrals correspond to $C_1,\dots,C_5$, respectively. On C_1, for instance, dr and $d\theta$ are zero because r and θ are constant on C_1 so we omitted the $-r\cos\theta\, dr + r^3\, d\theta$ part of the integrand. ∎

EXAMPLE 7. *Spherical Coordinates.* Verify Stokes's theorem for the case where

$$\mathbf{v} = \rho\hat{\mathbf{e}}_\rho - \rho^2\hat{\mathbf{e}}_\theta, \tag{41}$$

where $\mathcal{S}$ is as shown in Fig. 12 (a quarter of a cone, of semi-angle $\pi/6$), and where C is oriented as shown in the figure.

From the spherical coordinate expression for $\nabla \times \mathbf{v}$ in Section 16.7,

$$\nabla \times \mathbf{v} = -\rho\cot\phi\hat{\mathbf{e}}_\rho + 3\rho\hat{\mathbf{e}}_\phi. \tag{42}$$

Further, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_\phi$ and $dA = \rho|\sin\phi|\, d\rho\, d\theta$ since $\mathcal{S}$ is a constant-ϕ surface so

$$\begin{aligned}\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\, dA &= \int_0^{\pi/2}\int_0^2 \hat{\mathbf{e}}_\phi\cdot(-\rho\cot\phi\hat{\mathbf{e}}_\rho + 3\rho\hat{\mathbf{e}}_\phi)\,(\rho|\sin\phi|\, d\rho\, d\theta)\Big|_{\phi=\pi/6} \\ &= \frac{3}{2}\int_0^{\pi/2}\int_0^2 \rho^2\, d\rho\, d\theta = 2\pi. \end{aligned}\tag{43}$$

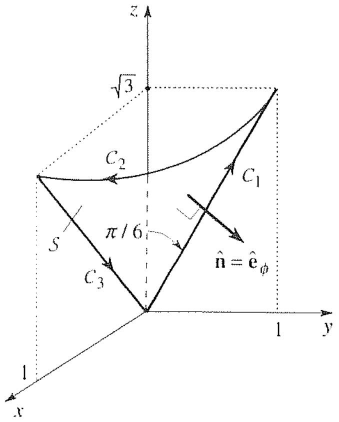


Figure 12. $\mathcal{S}$ and C for Example 7.

Next,

$$\begin{aligned}\oint_C \mathbf{v}\cdot d\mathbf{R} &= \oint_C (\rho\hat{\mathbf{e}}_\rho - \rho^2\hat{\mathbf{e}}_\theta)\cdot(d\rho\,\hat{\mathbf{e}}_\rho + \rho\, d\phi\,\hat{\mathbf{e}}_\phi + \rho|\sin\phi|\, d\theta\hat{\mathbf{e}}_\theta) \\ &= \oint_C \rho\, d\rho - \rho^3\sin\phi\, d\theta \\ &= \int_0^2 \rho\, d\rho\Big|_{\phi=\pi/6,\theta=\pi/2} + \int_{\pi/2}^0 (-\rho^3\sin\phi)\, d\theta\Big|_{\rho=2,\phi=\pi/6} + \int_2^0 \rho\, d\rho\Big|_{\phi=\pi/6,\theta=0} \\ &= 2 + 2\pi - 2 = 2\pi, \end{aligned}\tag{44}$$

in agreement with (43).

COMMENT. If the expressions used for dA and $d\mathbf{R}$ are not clear to you, see equation (18) in Section 15.6. If $\mathcal{S}$ were not a constant-coordinate surface we could not use (18) for dA, but we can always fall back on the $dA = \sqrt{EG - F^2}\, du\, dv$ formula in Section 15.5. Similarly, in this example $\hat{\mathbf{n}} = \hat{\mathbf{e}}_\phi$ is clear by inspection, but in more difficult cases we can always fall back on the gradient method [see equation (20) in section 16.8]. ∎

Closure. Having already studied line integrals in the form $\int_C f\, ds$, we first discuss line integrals that occur in the form $\int_C \mathbf{v}\cdot d\mathbf{R}$ and note their equivalance, except for the fact that C is not oriented in the former but it is in the latter. The latter form is prominent in Stokes's theorem,

$$\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\, dA = \oint_C \mathbf{v}\cdot d\mathbf{R}, \tag{45}$$

which evaluates the surface integral on $\mathcal{S}$ in terms of an integral along its boundary curve C. Remember that $\mathcal{S}$ in (45) is an open surface, not a closed surface, and that $\hat{\mathbf{n}}$ is not the "outward" normal because inward and outward are not meaningful for an open surface. Rather, the field of $\hat{\mathbf{n}}$'s on $\mathcal{S}$ are related to the orientation of $\mathcal{C}$ according to the right-hand rule. Finally, we derive Green's theorem (33), and observe that it is actually just a special case of Stokes's theorem for the case where $\mathbf{v} = P(x, y)\hat{\mathbf{i}} + Q(x, y)\hat{\mathbf{j}}$ and where $\mathcal{S}$ is a flat surface in the x, y plane.

Not only are the fundamental theorem of the integral calculus, Stokes's theorem, and the divergence theorem similar in expressing an integral over an n-dimensional region in terms of an evaluation over its $(n-1)$-dimensional boundary, they can all be generalized to a single statement in n dimensions using an elegant theory known as the theory of **differential forms**.*

EXERCISES 16.9

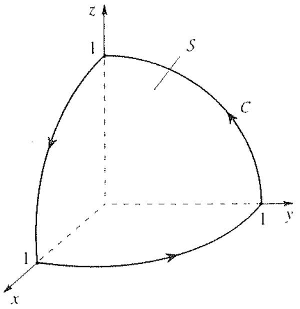
1. Derive the result $\int_{C_3} \mathbf{v} \cdot d\mathbf{R} = -\frac{259}{6}$, claimed in the paragraph following Example 1.

2. Verify Stokes's theorem. $\mathcal{S}$ is a plane surface with straight edges. The vertices and orientation of $\mathcal{C}$ are given.

(a) $\mathbf{v} = xy\hat{\mathbf{i}} - (2x - y)\hat{\mathbf{k}}$, $\mathcal{C}: (0,0,0)$ to $(1,1,0)$ to $(1,0,0)$ to $(0,0,0)$
(b) $\mathbf{v} = xy^2\hat{\mathbf{i}} - y^3\hat{\mathbf{j}} + 4xyz\hat{\mathbf{k}}$, $\mathcal{C}: (0,1,0)$ to $(1,0,0)$ to $(1,1,0)$ to $(0,1,0)$
(c) $\mathbf{v} = y^2\hat{\mathbf{i}} - x^2\hat{\mathbf{j}} - \sin(x^2y^2z^2)\hat{\mathbf{k}}$, $\mathcal{C}: (0,0,0)$ to $(2,0,0)$ to $(1,-1,0)$ to $(0,0,0)$
(d) $\mathbf{v} = xe^y\hat{\mathbf{i}} + (x + z)\hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathcal{C}: (-1,0,0)$ to $(0,1,0)$ to $(0,-1,0)$ to $(-1,0,0)$
(e) $\mathbf{v} = x^2yz\hat{\mathbf{j}}$, $\mathcal{C}: (0,1,0)$ to $(1,1,0)$ to $(1,0,1)$ to $(0,0,1)$ to $(0,1,0)$
(f) $\mathbf{v} = z\hat{\mathbf{i}} + y\hat{\mathbf{j}} + x\hat{\mathbf{k}}$, $\mathcal{C}: (1,0,0)$ to $(0,1,0)$ to $(-1,0,0)$ to $(0,-1,0)$ to $(1,0,0)$
(g) $\mathbf{v} = xy^2z\hat{\mathbf{j}}$, $\mathcal{C}: (0,0,1)$ to $(0,1,2)$ to $(1,2,0)$ to $(0,0,1)$
(h) $\mathbf{v} = xyz\hat{\mathbf{j}}$, $\mathcal{C}: (0,-1,0)$ to $(0,0,2)$ to $(1,1,0)$ to $(0,-1,0)$
(i) $\mathbf{v} = x^2z\hat{\mathbf{k}}$, $\mathcal{C}: (1,-1,0)$ to $(1,1,0)$ to $(0,0,1)$ to $(1,-1,0)$

3. Verify Stokes's theorem.

(a) $\mathbf{v} = x^2\hat{\mathbf{i}} + xz\hat{\mathbf{j}}$. $\mathcal{S}$ is a plane surface with edge curve $\mathcal{C}$ as follows: straight line from $(0,0,1)$ to $(0,1,1)$, then a straight line from $(0,1,1)$ to $(1,1,0)$, then back to $(0,0,1)$ along a curve parametrized by $x = \tau^2$, $y = \tau$, $z = 1 - \tau^2$, where $\tau: 1 \to 0$.
(b) $\mathbf{v} = x^2\hat{\mathbf{i}} - xz\hat{\mathbf{j}}$. $\mathcal{S}$ is a plane surface with edge curve $\mathcal{C}$ as follows: straight line from $(1,2,0)$ to $(1,1,0)$, then along a curve parametrized by $x = \tau^2$, $y = \tau$, $z = 1 - \tau^2$, where $\tau: 1 \to 0$, then along a curve parametrized by $x = \tau^2$, $y = 2\tau$, $z = 1 - \tau^2$, where $\tau: 0 \to 1$.
(c) $\mathbf{v} = y^2\hat{\mathbf{j}} - xy^2z\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface of a unit cube with corners at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,0,1)$, $(1,1,0)$, $(0,1,1)$, $(1,1,1)$, excluding the face $x = 1$. Take $\hat{\mathbf{n}} = \hat{\mathbf{k}}$ on the $z = 1$ face.
(d) $\mathbf{v} = -y\hat{\mathbf{i}} + x\hat{\mathbf{j}} + 3\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $z = 4 - x^2 - y^2$ between $z = 0$ and $z = 4$, and $\hat{\mathbf{n}} = \hat{\mathbf{k}}$ at the point $(0,0,4)$ on $\mathcal{S}$.
(e) $\mathbf{v} = x^2y\hat{\mathbf{j}}$. $\mathcal{S}$ lies in the x, y plane and has edge curve $\mathcal{C}$ as follows: straight line from $(0,0,0)$ to $(0,-1,0)$, then a straight line from $(0,-1,0)$ to $(1,1,0)$, then along $y = x^2$ back to $(0,0,0)$.
(f) $\mathbf{v} = x\hat{\mathbf{i}} - xz\hat{\mathbf{k}}$. $\mathcal{S}$ lies in the x, z plane and has edge curve $\mathcal{C}$ as follows: $z = x$ from $(0,0,0)$ to $(1,0,1)$, then $x = z^2$ from $(1,0,1)$ to $(0,0,0)$.
(g) $\mathbf{v} = yz\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $z = 1 - x^2 - y^2$ cut off by the planes $x = 0$, $y = 0$, and $z = 0$, and $\mathcal{C}$ oriented as shown.

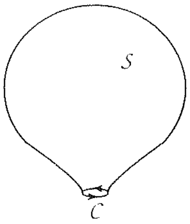


[*] For a readable introduction to differential forms, see Section 7.6 of J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Fransisco: W. H. Freeman, 1976).

(h) $\mathbf{v} = x^2z\hat{\mathbf{i}} - 3xy\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $x = 1 - z^2$ for $-1 \leq x \leq 1, 0 \leq y \leq 3$, with $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ at the point $(1, 1, 0)$ on $\mathcal{S}$.

(i) $\mathbf{v} = yz\hat{\mathbf{i}} + xy\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $x = z^2$ for $0 \leq x \leq 1$, $0 \leq y \leq 2$, with $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ at the point $(0, 1, 0)$ on $\mathcal{S}$.

4. Remember that $\mathcal{S}$ in Stokes's theorem is an open surface. Suppose we let $\mathcal{S}$ tend to a closed surface by letting $\mathcal{C}$ get very small (see the figure), until it shrinks to a point. Since $\mathcal{C}$ shrinks to a point, $\oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} \to 0$ and, by Stokes's theorem,



$$\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA \to 0 \tag{4.1}$$

as well. Show that the result (4.1) also follows from the divergence theorem.

5. The following claim was put forth in an examination paper: "The line integral $\oint_{\mathcal{C}} f(x, y)\, dx + g(x, y)\, dy$, where $\mathcal{C}$ is the clockwise unit circle, is necessarily zero because on $\mathcal{C}$ we can express y as a function of x so that [taking $(1, 0)$ as the initial point of $\mathcal{C}$ and also the terminal point of $\mathcal{C}$]

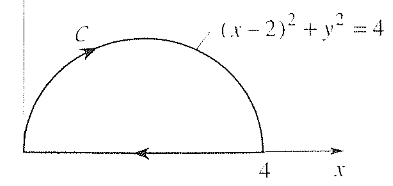
$$\oint_{\mathcal{C}} f(x, y)\, dx = \int_1^1 f(x, y(x))\, dx = \int_1^1 F(x)\, dx = 0$$

and, similarly,

$$\oint_{\mathcal{C}} g(x, y)\, dy = \int_0^0 g(x(y), y)\, dy = \int_0^0 G(y)\, dy = 0."$$

Give a critical assessment of that claim. Is it true? False? Explain your reasoning.

6. Given the field $\mathbf{F} = y^2\hat{\mathbf{i}}$ and the contour shown, the following conflicting calculations were put forward in an exam paper:



$$\oint_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{R} = \oint_{\mathcal{C}} y^2\, dx$$
$$= \int_0^4 [4 - (x-2)^2]\, dx + \int_4^0 0\, dx = \frac{32}{3},$$

but

$$\oint_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{R} = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{F}\, dA$$
$$= \int_{\mathcal{S}} 2y\, dA = 2\int_0^4 y(y\, dx)$$
$$= 2\int_0^4 [4 - (x-2)^2]\, dx = \frac{64}{3}.$$

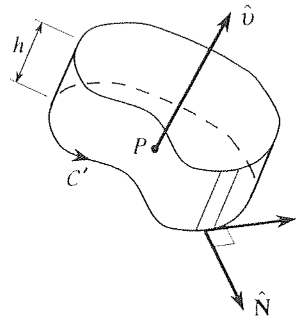
Find, and correct, the error.

7. (*An observation about Stokes's theorem*) Notice that the value of the line integral in Stokes's theorem (18) is independent of the shape of the surface $\mathcal{S}$. Thus, the surface integral in (18) must, similarly, be independent of the shape $\mathcal{S}$. That is, it must be true that $\int_{\mathcal{S}_1} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_{\mathcal{S}_2} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA$ for any two surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ that share the same edge curve $\mathcal{C}$. Provide an alternative argument as to why the two integrals (in the preceding sentence) must be equal. HINT: Note that $\mathcal{S}_1 + \mathcal{S}_2$ is a closed surface. Apply the divergence theorem to the surface integral of $\hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}$ over that closed surface.

8. (*Heuristic proof of Stokes's theorem*) Our proof of Stokes's theorem was limited to the case where $\mathcal{S}$ is flat. This exercise is to indicate how to prove the theorem for a general surface $\mathcal{S}$ that is not necessarily flat. We use the limit definition of the curl,

$$\operatorname{curl} \mathbf{v}(P) \equiv \lim_{\mathcal{B}\to 0} \left\{ \frac{\int_{\mathcal{S}'} \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V} \right\}, \tag{8.1}$$

given in Exercise 6 of Section 16.5. Let $\mathcal{B}$ be a right circular cylinder of a small height h, and let the base of $\mathcal{B}$ be a plane region of area A, bounded by a simple closed curve $\mathcal{C}'$, oriented so that the unit normal $\hat{\nu}$ to the base, and $\mathcal{C}'$, are in accordance with the right-hand rule as seen in the figure.



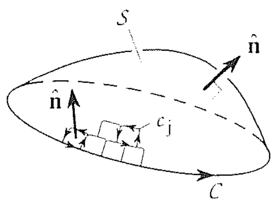
Dotting both sides of (8.1) with $\hat{\nu}$, and recalling that $\mathbf{A}\cdot\mathbf{B}\times\mathbf{C} = \mathbf{A}\times\mathbf{B}\cdot\mathbf{C}$, show that

$$\hat{\nu}\cdot\operatorname{curl}\mathbf{v} = \lim_{A\to 0}\left\{\frac{\int_{C'}\mathbf{v}\cdot d\mathbf{R}}{A}\right\}, \tag{8.2}$$

or, in differential form,

$$\hat{\nu}\cdot\operatorname{curl}\mathbf{v}\,dA \sim \int_{C'}\mathbf{v}\cdot d\mathbf{R} \tag{8.3}$$

as $A \to 0$. Now take the surface S in Stokes's theorem and partition it with a large number of such curves as C', as suggested in the figure, and number them as $C_1', C_2', \ldots$. Write down (8.3) for each curve $(C_1', C_2', \ldots)$, realizing that $\hat{\nu}$ is the normal $\hat{\mathbf{n}}$ to S. Add these equations and, noting internal cancellation analogous to that which occured in our proof of the divergence theorem, show that Stokes's theorem follows. [Note that the S' in (8.1) is the closed surface of the infinitesimal body $\mathcal{B}$, whereas the S in the figure above, is the open cap-like surface in Stokes's theorem.]



9. (*Faraday's law*) **Faraday's law** (that the emf around a closed curve equals the negative of the time rate of change of the magnetic flux through the curve) may be expressed as

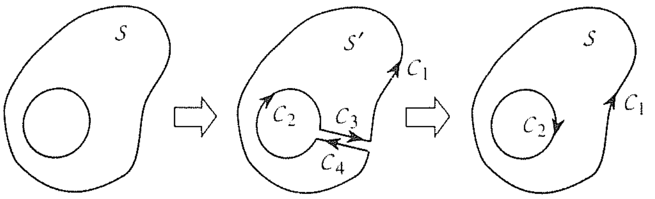
$$\begin{aligned}\oint_C \mathbf{E}\cdot d\mathbf{R} &= -\frac{d}{dt}\int_S \mathbf{B}\cdot\hat{\mathbf{n}}\,dA \\ &= -\int_S \frac{\partial\mathbf{B}}{\partial t}\cdot\hat{\mathbf{n}}\,dA\end{aligned} \tag{9.1}$$

for every (fixed) surface S with closed boundary curve C in the field, where $\mathbf{E}$ is the electric field intensity, $\mathbf{B}$ is the magnetic flux density, t is the time, and where the relative orientations of C and $\hat{\mathbf{n}}$ are the same as in Stokes's theorem. Applying Stokes's theorem to the line integral, use the arbitrariness of S and C to deduce (heuristically) that the relation

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \tag{9.2}$$

holds at all points in the field. Equation (9.2) is one of the *Maxwell's equations* for time-varying fields; for steady fields it reduces to $\nabla\times\mathbf{E} = 0$.

10. (*Surfaces with holes*) Here we indicate a modest extension of Stokes's theorem to the case where S has one or more holes in it. It should suffice to consider the surface S shown in the left-hand figure which, for simplicity, has only one hole in it and lies in the plane of the paper. Thus, the boundary C of S is not a simple closed curve, as is called for in Stokes's theorem.



(a) Slit the region as shown in the middle figure, and define the contours C_1, C_2, C_3, C_4 enclosing the region S'. Since S' and $C = C_1 + C_2 + C_3 + C_4$ do satisfy the requirements given in Stokes's theorem, that theorem can now be applied. Letting the gap between C_3 and C_4 tend to zero, show that we obtain

$$\boxed{\int_S \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA = \oint_{C_1}\mathbf{v}\cdot d\mathbf{R} + \oint_{C_2}\mathbf{v}\cdot d\mathbf{R},} \tag{10.1}$$

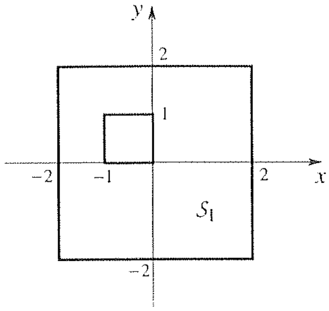
in accordance with the right-hand figure, where $\hat{\mathbf{n}}$ is directed out of the paper, toward the reader. NOTE: Since Green's theorem is but a special case of Stokes's theorem, a result analogous to (10.1) follows immediately for Green's theorem:

$$\boxed{\begin{aligned}\int_S\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dA = &\oint_{C_1} P\,dx + Q\,dy \\ &+ \oint_{C_2} P\,dx + Q\,dy.\end{aligned}} \tag{10.2}$$

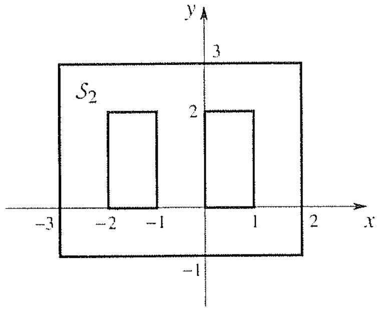
(b) Using pictures and reasoning, derive a formula analogous to (10.1) for the case where S has two holes rather than one.

11. Verify Stokes's theorem (10.1) above.

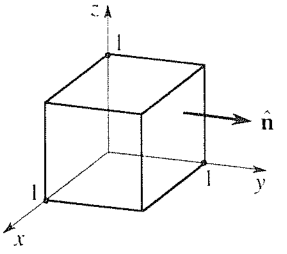
(a) $\mathbf{v} = y^3\hat{\mathbf{i}}$, S is the region S_1 shown, and $\hat{\mathbf{n}} = \hat{\mathbf{k}}$.



(b) $\mathbf{v} = xy\hat{\mathbf{i}}$, S is the region S_2 shown, and $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$.



(c) $\mathbf{v} = x^3y(1+z)\hat{\mathbf{i}} + 3\cos(x^2+z^2)\hat{\mathbf{k}}$, and S is the four-sided surface shown, open at the top and bottom.

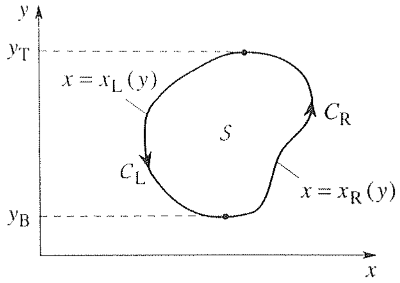


12. (*Direct proof of Green's theorem*) Green's theorem (33) can also be proved by treating the left-hand side as an iterated integral and integrating directly. Suppose that S is convex in both x and y. By **convex in** x, for instance, we mean that S can be described by the statement

$$x_L(y) \le x \le x_R(y) \quad \text{for} \quad y_B \le y \le y_T. \qquad (12.1)$$

That is, every horizontal line from C_L to C_R (see figure) lies entirely within S. Then

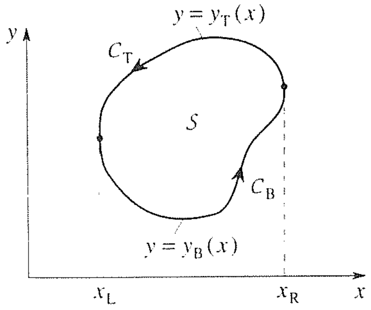
$$\int_S \frac{\partial Q}{\partial x}\,dA = \int_{y_B}^{y_T}\int_{x_L(y)}^{x_R(y)} \frac{\partial Q}{\partial x}\,dx\,dy = \cdots = \oint_C Q\,dy. \qquad (12.2)$$



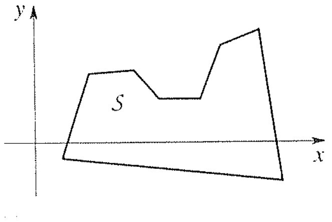
(a) Fill in the missing steps in (12.2).
(b) Next, show that

$$\int_S \frac{\partial P}{\partial y}\,dA = \int_{x_L}^{x_R}\int_{y_B(x)}^{y_T(x)} \frac{\partial P}{\partial y}\,dy\,dx = \cdots = -\oint_C P\,dx, \qquad (12.3)$$

where $x_L, x_R, y_B(x)$ and $y_T(x)$ are shown below; (12.2) and (12.3) then give (33).



(c) Indicate how you could extend the convex-S result, proved above, to cover nonconvex regions such as the one shown below.



13. If C is a piecewise smooth simple closed curve in the x, y plane, oriented counterclockwise, show that the area A enclosed by C is given by the line integral

$$A = \frac{1}{2}\oint_C (x\,dy - y\,dx). \qquad (13.1)$$

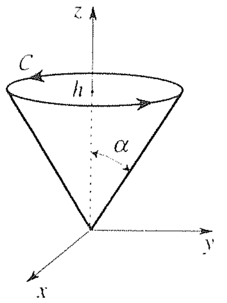
14. Use (13.1) in Exercise 13 to compute A if C is the boundary of the

(a) rectangle with vertices at $(0,0)$, $(a,0)$, (a,b), $(0,b)$
(b) triangle with vertices at $(a,0)$, (a,b), $(0,b)$
(c) circle $x^2+y^2=9$

15. Show that (33) does *not* hold for the case where $P = -y/(x^2+y^2)$, $Q = x/(x^2+y^2)$, and $\mathcal{S}$ is the unit disk $x^2+y^2 \leq 1$. Explain why this failure does not violate Green's theorem.

16. Verify Stokes's theorem.

(a) $\mathbf{v} = \omega r \hat{\mathbf{e}}_\theta$ ($\omega =$ constant), $\mathcal{S}$: the waterglass surface consisting of the cylinder $r = a$ over $0 \leq z \leq h$, with flat base $z = 0$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the base.
(b) $\mathbf{v} = 3r\hat{\mathbf{e}}_r - rz^2\hat{\mathbf{e}}_\theta + 5r^2\hat{\mathbf{e}}_z$, $\mathcal{S}$: the upside-down waterglass surface consisting of the cylinder $r = a$ over $0 \leq z \leq h$, with flat top $z = h$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the face $z = h$.
(c) $\mathbf{v} = (50 - 2y)\hat{\mathbf{i}}$, $\mathcal{S}$: the flat disk $r \leq 1$, $z = 0$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$.
(d) $\mathbf{v} = \omega r \hat{\mathbf{e}}_\theta$ ($\omega =$ constant), $\mathcal{S}$: the conical surface, shown below, open at the top, C oriented as shown.



(e) Repeat part (d), but with $\mathbf{v} = 3r\hat{\mathbf{e}}_r - r^4\hat{\mathbf{e}}_\theta$.
(f) Repeat part (d), but with $\mathbf{v} = \rho^2 \sin\phi \hat{\mathbf{e}}_\theta$.

16.10 Irrotational Fields

16.10.1. Irrotational fields. Let $\mathbf{v}$ be a C^1 vector field in a domain $\mathcal{D}$.[*] If

$$\boxed{\nabla \times \mathbf{v} = \mathbf{0}} \tag{1}$$

at each point in $\mathcal{D}$, then $\mathbf{v}$ is said to be **irrotational** in $\mathcal{D}$. In Section 16.5 we saw that $\nabla \times \mathbf{v}$ can be interpreted physically as twice the fluid particle angular velocity at the point in question. Thus, if $\mathbf{v}$ is irrotational that means that the particles have no angular velocity, or "spin." For instance, the plane flow over a semicircular bump given by equation (3) in Section 16.2, and shown there in Fig. 4, was irrotational because $\nabla \times \mathbf{v} = \mathbf{0}$. In fact, irrotational flows form an important case in applications such as aerodynamics, water wave mechanics, and gravitational fields.

To study irrotational fields we need to define two more terms. First, we say that the line integral $\int_C \mathbf{v} \cdot d\mathbf{R}$ between endpoints P and Q, is **path independent** if the value of the integral is the same for *every* piecewise smooth path C lying within $\mathcal{D}$; if it is path independent for *every* pair of endpoints P and Q in $\mathcal{D}$ then we say that it is path independent in $\mathcal{D}$.

Second, we need to add to our topological concept of connectedness, which was introduced in Section 13.2.2. A domain $\mathcal{D}$ is said to be **simply connected** if every closed curve in $\mathcal{D}$ can be shrunk, by a continuous deformation, to any point in $\mathcal{D}$. If it is not simply connected it is **multiply connected**. For instance, all of 3-space, the interior of a sphere or cube and the region between two concentric

[*] Recall from Section 13.2.2 that a domain is an open region, i.e., a connected set containing none of its boundary points.

spherical surfaces are simply connected, but the interior of a torus (i.e., a doughnut, bagel, or wedding band) is not; it is multiply connected. That is, whereas a closed curve such as C_1 (Fig. 1) can be shrunk to any desired point within $\mathcal{D}$, a closed curve such as C_2 cannot. In two dimensions, all of 2-space and the interior of a circle or square are simply connected, but the region between two concentric circles is not.

The focus of this section is the following theorem and its applications.

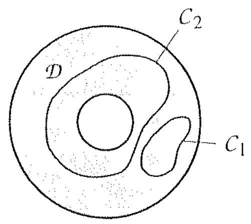


Figure 1. Torus.

THEOREM 16.10.1 *Irrotational Field*
If $\mathbf{v}$ is C^1 in a simply connected domain $\mathcal{D}$, then the following statements are equivalent (i.e., if one holds, then the other three do also).

(a) There exists a C^2 scalar function Φ in $\mathcal{D}$ such that $\mathbf{v} = \boldsymbol{\nabla}\Phi$ throughout $\mathcal{D}$.

(b) $\boldsymbol{\nabla} \times \mathbf{v} = \mathbf{0}$ throughout $\mathcal{D}$.

(c) $\int_C \mathbf{v} \cdot d\mathbf{R} = 0$ for every piecewise smooth simple closed path C in $\mathcal{D}$.

(d) $\int_C \mathbf{v} \cdot d\mathbf{R}$ is independent of path in $\mathcal{D}$.

Proof: Because of the claimed equivalence one might believe that we need to prove that (a) $\Rightarrow$ (b), (a) $\Rightarrow$ (c), (a) $\Rightarrow$ (d), (b) $\Rightarrow$ (a), (b) $\Rightarrow$ (c), (b) $\Rightarrow$ (d), and so on, twelve items altogether. No, it suffices to prove any closed logical loop, such as (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a).

(a) $\Rightarrow$ (b): If $\mathbf{v} = \boldsymbol{\nabla}\Phi = \Phi_x\hat{\mathbf{i}} + \Phi_y\hat{\mathbf{j}} + \Phi_z\hat{\mathbf{k}}$ (with subscripts denoting partial derivatives, as can be understood from the context), then $\boldsymbol{\nabla} \times \mathbf{v} = (\Phi_{zy} - \Phi_{yz})\hat{\mathbf{i}} - (\Phi_{zx} - \Phi_{xz})\hat{\mathbf{j}} + (\Phi_{yx} - \Phi_{xy})\hat{\mathbf{k}} = \mathbf{0}$ because it follows from the assumption that Φ is C^2 that $\Phi_{yz} = \Phi_{zy}$, $\Phi_{zx} = \Phi_{xz}$, and $\Phi_{yx} = \Phi_{xy}$ (Theorem 13.3.1). Or, remember from (13) in Section 16.6 that the curl of a gradient is zero.

(b) $\Rightarrow$ (c): Let C be any piecewise smooth simple closed path within $\mathcal{D}$. Suppose that we can introduce a piecewise smooth surface $\mathcal{S}$ with C as its boundary. (Note that if $\mathcal{D}$ were not simply connected, $\mathcal{S}$ might have one or more holes in it, and then C would be only part of the boundary of $\mathcal{S}$. See Exercise 9 of Section 16.9.) Then, by Stokes's theorem,

$$\oint_C \mathbf{v} \cdot d\mathbf{R} = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \boldsymbol{\nabla} \times \mathbf{v}\, dA = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{0}\, dA = 0, \tag{2}$$

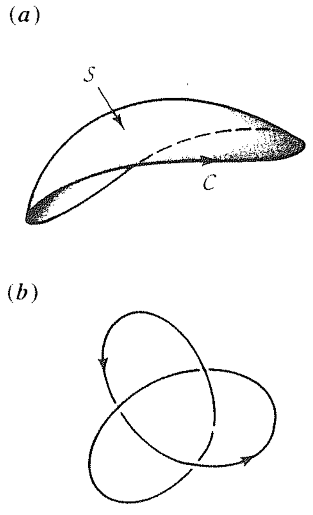


Figure 2. Is there an $\mathcal{S}$?

as claimed. However, given a piecewise smooth simple closed path C, can we necessarily find a piecewise smooth surface $\mathcal{S}$ having C as its boundary, as assumed above? For the C shown in Fig. 2a, for example, such an $\mathcal{S}$ is readily constructed as sketched in the figure. But what about more complicated C's – for example, the overhand knot shown in Fig. 2a. In that case the idea is to add cancelling line segments, shown as dashed lines in Fig. 3, so that the curve C can be split into two noninterlocking curves C_1 and C_2 which *do* admit suitable surfaces S_1 and S_2, respectively. Then, analogous to (1) we have

$$\oint_C \mathbf{v}\cdot d\mathbf{R} = \oint_{C_1} \mathbf{v}\cdot d\mathbf{R} + \oint_{C_2} \mathbf{v}\cdot d\mathbf{R} = \int_{S_1} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA + \int_{S_2} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA$$
$$= \int_{S_1} \hat{\mathbf{n}}\cdot\mathbf{0}\,dA + \int_{S_2} \hat{\mathbf{n}}\cdot\mathbf{0}\,dA = 0 + 0 = 0.$$

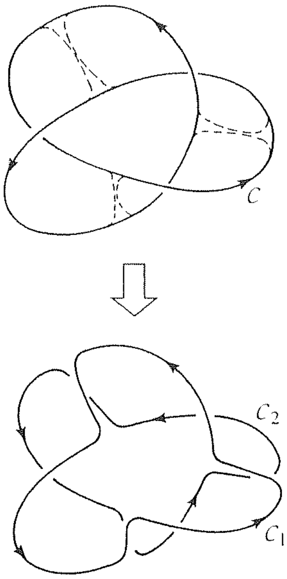


Figure 3. Undoing the knot.

Rigorous proof that we can handle *any* piecewise smooth simple closed path C, however, is beyond our present scope.

(c) ⇒ (d): Let C_1 and C_2 be any two piecewise smooth paths from any point $P_0 = (x_0, y_0, z_0)$ in $\mathcal{D}$ to any point $P = (x, y, z)$ in $\mathcal{D}$ (Fig. 4a). Suppose that C_1 and C_2 are simple curves and that they intersect each other only at the endpoints P_0 and P. (The argument needed for the case where additional intersections occur will be omitted here.) Then $C_1 + (-C_2)$ is a piecewise smooth simple closed path (Fig. 4b) and, according to item (c),

$$\oint_{C_1+(-C_2)} \mathbf{v}\cdot d\mathbf{R} = 0. \tag{3}$$

But

$$\oint_{C_1+(-C_2)} \mathbf{v}\cdot d\mathbf{R} = \oint_{C_1} \mathbf{v}\cdot d\mathbf{R} + \int_{-C_2} \mathbf{v}\cdot d\mathbf{R}$$
$$= \int_{C_1} \mathbf{v}\cdot d\mathbf{R} - \int_{C_2} \mathbf{v}\cdot d\mathbf{R}, \tag{4}$$

and it follows from (3) and (4) that

$$\int_{C_1} \mathbf{v}\cdot d\mathbf{R} = \int_{C_2} \mathbf{v}\cdot d\mathbf{R}, \tag{5}$$

as claimed in (d).

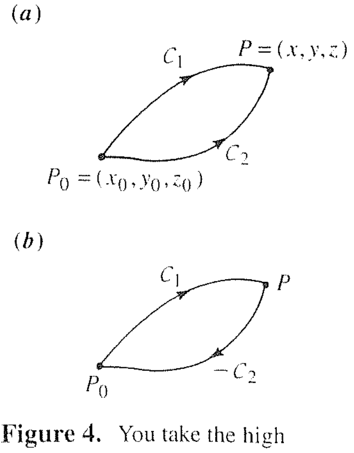


Figure 4. You take the high road, I'll take the low.

(d) ⇒ (a): Regarding $P_0 = (x_0, y_0, z_0)$ as fixed and $P = (x, y, z)$ as a variable endpoint, it follows from the path independence [item (d)] that

$$f(x, y, z) \equiv \int_{(x_0,y_0,z_0)}^{(x,y,z)} \mathbf{v}\cdot d\mathbf{R} \tag{6}$$

is indeed a function of x, y, z. For if different paths to (x, y, z) were to give different values of the integral, then the integral would not be uniquely determined by x, y, z. That is, it would not be a single-valued function of x, y, z and, since functions are to be single-valued, it would not be a function of x, y, z.

Consider first the x dependence of the integral. Regarding y and z as fixed, take the path from (x_0, y_0, z_0) to (x, y, z) in (6) to be the one shown in Fig. 5. That is, the path is fixed, except that the endpoint $P = (x, y, z)$ moves parallel to the x axis as x is varied, with y and z fixed. Then (6) becomes

$$\begin{aligned} f(x,y,z) &= \int_{(x_0,y_0,z_0)}^{(x_0,y,z)} \mathbf{v}\cdot d\mathbf{R} + \int_{(x_0,y,z)}^{(x,y,z)} \mathbf{v}\cdot d\mathbf{R} \\ &= \text{constant} + \int_{(x_0,y,z)}^{(x,y,z)} (v_x\,dx + v_y\,dy + v_z\,dz) \\ &= \text{constant} + \int_{x_0}^{x} v_x(x,y,z)\,dx, \end{aligned} \tag{7}$$

where $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$, and where the last step follows from the fact that y and z are constant along the path from (x_0, y, z) to (x, y, z). Or, introducing a dummy variable of integration to avoid confusion, we have

$$f(x,y,z) = \text{constant} + \int_{x_0}^{x} v_x(\xi, y, z)\,d\xi, \tag{8}$$

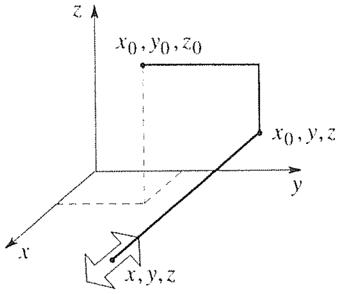


Figure 5. The x dependence of f.

and since v_x is assumed to be continuous it follows from the fundamental theorem of the integral calculus that $\partial f/\partial x = v_x(x, y, z)$. In the same manner, we may show that $\partial f/\partial y = v_y(x, y, z)$, and that $\partial f/\partial z = v_z(x, y, z)$. Thus, $\mathbf{v} = \nabla f$ and f is the scalar function that we've been looking for. ■

With Theorem 16.10.1 in hand, suppose we wish to evaluate a given line integral

$$I = \int_C \mathbf{v}\cdot d\mathbf{R}. \tag{9}$$

First, check to see if $\mathbf{v}$ is irrotational (i.e., if $\nabla \times \mathbf{v} = \mathbf{0}$) in a simply connected domain $\mathcal{D}$ containing the path C. If it is, and C is a closed path (piecewise smooth simple closed path, to be precise), then it follows immediately from item (c) that $I = 0$. If $\mathbf{v}$ is irrotational but C is not closed, then I is not necessarily zero, yet Theorem 16.10.1 still provides two alternative simplifications. First, we can use (d) to justify changing the path to a simpler one, keeping the endpoints fixed. Or we can use (a), which guarantees the existence of a scalar function (i.e., a scalar field) $\Phi(x, y, z)$ such that $\nabla\Phi = \mathbf{v}$. For if we can solve the relation $\nabla\Phi = \mathbf{v}$ for Φ, which function is known as the **scalar potential** or simply the **potential**, corresponding to the vector field $\mathbf{v}$, then

$$\begin{aligned} I = \int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} &= \int_{P_i}^{P_f} \nabla\Phi\cdot d\mathbf{R} \\ &= \int_{P_i}^{P_f} \left(\frac{\partial\Phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\Phi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\Phi}{\partial z}\hat{\mathbf{k}}\right)\cdot\left(dx\,\hat{\mathbf{i}} + dy\,\hat{\mathbf{j}} + dz\,\hat{\mathbf{k}}\right) \\ &= \int_{P_i}^{P_f} \left(\frac{\partial\Phi}{\partial x}dx + \frac{\partial\Phi}{\partial y}dy + \frac{\partial\Phi}{\partial z}dz\right) = \int_{P_i}^{P_f} d\Phi = \Phi\Big|_{P_i}^{P_f}, \end{aligned} \tag{10}$$

where P_i is the initial point and P_f is the final point. That is,

$$\boxed{\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f},} \tag{11}$$

the line integral is equal to the change in the potential.

Both of these lines of approach are illustrated in the example to follow.

EXAMPLE 1. Evaluate the line integral $I = \int_C \mathbf{v}\cdot d\mathbf{R}$, where

$$\mathbf{v} = 3x^2y^2\hat{\mathbf{i}} + (2x^3y - e^z)\hat{\mathbf{j}} + (2z - ye^z)\hat{\mathbf{k}}, \tag{12}$$

and where C is given parametrically by

$$\begin{aligned} x &= \left(1 - \frac{\tau}{2}\right)e^{\sqrt{\tau}} - 2\frac{\sin\tau^3}{\sin 8}\sqrt[3]{\frac{\tau}{2}}, \\ y &= -\frac{6}{\tau+3} + \frac{21}{10}\tau, \\ z &= \frac{1}{16}\tau^5 - 1 \end{aligned} \tag{13}$$

for $\tau : 0 \to 2$. We could, of course, put the parametrization (13) into

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \int_C 3x^2y^2\,dx + (2x^3y - e^z)\,dy + (2z - ye^z)\,dz. \tag{14}$$

That step would give an "ordinary" integral of the form $\int_0^2 F(\tau)\,d\tau$. However, the integrand $F(\tau)$ would, evidently, be quite unwieldy. Thus, it is best to check, first, to see if $\mathbf{v}$ is irrotational. In fact, we do find that $\nabla \times \mathbf{v} = \mathbf{0}$ so we can use either of the methods outlined above. Thus, we can discard the unwieldy path (13); all we need to extract from (13) are the initial and final points. Setting $\tau = 0$ and 2 in (13) gives $P_i = (1, -2, -1)$ and $P_f = (-2, 3, 1)$.

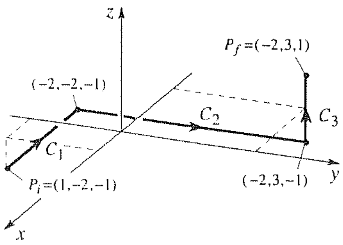


Figure 6. A simple path.

Let us evaluate I first by path simplification. What *is* a simple path? Surely, a straight line from P_i to P_f comes to mind. No, the simplest path will be one along which only one coordinate varies. Specifically, let us use the path C_1, C_2, C_3 shown in Fig. 6. Then

$$\begin{aligned} I &= \int_{C_1} \mathbf{v}\cdot d\mathbf{R} + \int_{C_2} \mathbf{v}\cdot d\mathbf{R} + \int_{C_3} \mathbf{v}\cdot d\mathbf{R} \\ &= \int_{C_1} 3x^2y^2\,dx\Big|_{y=-2,z=-1} + \int_{C_2} (2x^3y - e^z)\,dy\Big|_{x=-2,z=-1} \\ &\quad + \int_{C_3} (2z - ye^z)\,dz\Big|_{x=-2,y=3} \\ &= 12\int_1^{-2} x^2\,dx + \int_{-2}^{3} (-16y - e^{-1})\,dy + \int_{-1}^{1} (2z - 3e^z)\,dz \\ &= -76 - 3e - 2e^{-1}. \end{aligned} \tag{15}$$

To appreciate the simplification achieved, observe that in the C_1 integral we kept only the dx part because y and z are constant on C_1 (hence the dy's and dz's are zero). Further, the integrand of the dx integral simplifies because any y's or z's contained therein are constant on C_1. Similarly for the C_2 and C_3 integrals.

Alternatively, let us evaluate I using the potential Φ, the existence of which is assured by the fact that $\nabla \times \mathbf{v} = \mathbf{0}$. To find $\Phi(x, y, z)$ we use the connection $\mathbf{v} = \nabla\Phi$ between $\mathbf{v}$ and Φ:

$$\frac{\partial \Phi}{\partial x} = v_x = 3x^2y^2, \tag{16a}$$

$$\frac{\partial \Phi}{\partial y} = v_y = 2x^3y - e^z, \tag{16b}$$

$$\frac{\partial \Phi}{\partial z} = v_z = 2z - ye^z. \tag{16c}$$

We need to integrate (16a,b,c) to solve for Φ. First, integrate (16a) with respect to x, holding y and z fixed (since y and z were held fixed in the derivative $\partial\Phi/\partial x$) :

$$\Phi(x, y, z) = \int 3x^2y^2\, \partial x = x^3y^2 + A(y, z), \tag{17}$$

where the ∂x notation is not standard in the literature but may be helpful in reminding us that y and z are fixed. We need to allow the integration constant A to depend on y and z; observe that $\partial/\partial x$ of (17) does give us back (16a).

Next, put (17) into the left-hand side of (16b):

$$2x^3y + \frac{\partial A}{\partial y} = 2x^3y - e^z. \tag{18}$$

Cancelling the $2x^3y$ terms and integrating partially on y gives

$$A(y, z) = -\int e^z\, \partial y = -ye^z + B(z), \tag{19}$$

where we let the integration constant B depend on z because z was held fixed in the partial integration on y. Putting (19) into (17) gives the updated expression for Φ,

$$\Phi(x, y, z) = x^3y^2 - ye^z + B(z). \tag{20}$$

Finally, putting (20) into the left-hand side of (16c) gives

$$0 - ye^z + B'(z) = 2z - ye^z. \tag{21}$$

Cancelling the $-ye^z$ terms and integrating on z gives

$$B(z) = \int 2z\, dz = z^2 + C, \tag{22}$$

where we use dz rather than ∂z because it is an ordinary integral on z, and the integration constant C really is just a constant. Thus,

$$\Phi(x, y, z) = x^3y^2 - ye^z + z^2 + C. \tag{23}$$

Then (11) gives

$$\int_C \mathbf{v}\cdot d\mathbf{R} = (x^3y^2 - ye^z + z^2 + C)\Big|_{(1,-2,-1)}^{(-2,3,1)} = -76 - 3e - 2e^{-1}, \tag{24}$$

in agreement with (15).

COMMENT 1. There will always occur an arbitrary additive constant in Φ, just as C occurs in (23), for an additive constant drops out when we ask $\nabla\Phi$ to equal $\mathbf{v}$. However, when Φ is evaluated between the endpoints C inevitably cancels out so one can simply set $C = 0$, say, when it first appears.

COMMENT 2. One might wonder how the procedure of integrating $\nabla\Phi = \mathbf{v}$ can *fail* to produce a scalar potential $\Phi(x, y, z)$, whether or not $\mathbf{v}$ is irrotational. To understand this point, recall the cancellation of the $2x^3y$ terms. If those terms did not cancel, then (18) would have presented a logical contradiction for $\partial A/\partial y$ would therefore have depended on x, whereas it is a function only of y and z. Similarly, in (21) the cancellation of the $-ye^z$ terms was essential because otherwise $B'(z)$ would have depended on y. These cancellations were not accidents; they occurred because $\mathbf{v}$ was irrotational.

COMMENT 3. An advantage of the potential method over path simplification is that with $\Phi(x, y, z)$ in hand we can now compute $\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R}$ between *any* two points P_i and P_f, from (11). ∎

EXAMPLE 2. *Conservative Force Fields.* A force field $\mathbf{F}$ defined in a domain $\mathcal{D}$ is said to be **conservative** *if the work*

$$W = \oint_C \mathbf{F}\cdot d\mathbf{R} \tag{25}$$

is zero for every piecewise smooth simple closed path C *in* $\mathcal{D}$. If $\mathbf{F}$ is C^1 throughout $\mathcal{D}$, then it follows from Theorem 16.10.1 that $\mathbf{F}$ is conservative if and only if it is irrotational. If so, there exists a scalar potential Φ such that

$$\mathbf{F} = -\nabla\Phi, \tag{26}$$

where the minus sign is customary but not essential.

To illustrate, consider the uniform gravitational field $\mathbf{F} = -g\hat{\mathbf{k}}$ per unit mass, where g is the acceleration due to gravity. Surely $\nabla\times\mathbf{F} = \mathbf{0}$ since $\mathbf{F}$ is a constant. Then $\mathbf{F} = -g\hat{\mathbf{k}} = -\nabla\Phi$ so that

$$\frac{\partial\Phi}{\partial x} = 0, \qquad \frac{\partial\Phi}{\partial y} = 0, \qquad \frac{\partial\Phi}{\partial z} = g.$$

Integrating gives us

$$\Phi = gz + C, \tag{27}$$

which is the familiar "gravitational potential energy" from elementary physics.

Returning to the general case, suppose that $\mathbf{F}$ is conservative, and consider the work $\int \mathbf{F}\cdot d\mathbf{R}$ over any path from an initial point P_i to a final point P_f. With $\mathbf{F} = m\ddot{\mathbf{R}}$ from

Newton's second law, where m is the mass (assumed constant), we have

$$\begin{aligned}\int_{P_i}^{P_f} \mathbf{F}\cdot d\mathbf{R} &= \int_{P_i}^{P_f} m\ddot{\mathbf{R}}\cdot d\mathbf{R} = m\int \ddot{\mathbf{R}}\cdot\dot{\mathbf{R}}\,dt\\ &= \frac{1}{2}m\int \frac{d}{dt}(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}})\,dt = \frac{1}{2}m\int d(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}})\\ &= \frac{1}{2}m(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}})\Big|_{P_i}^{P_f} = \frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2\Big|_{P_i}^{P_f}. \end{aligned} \tag{28}$$

But since $\mathbf{F}$ is conservative, we may also express

$$\int_{P_i}^{P_f} \mathbf{F}\cdot d\mathbf{R} = \int_{P_i}^{P_f} -\nabla\Phi\cdot d\mathbf{R} = -\int_{P_i}^{P_f} d\Phi = -\Phi\Big|_{P_i}^{P_f}. \tag{29}$$

It follows from (28) and (29) that $\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2\Big|_{P_i}^{P_f} = -\Phi\Big|_{P_i}^{P_f}$, or

$$\left(\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi\right)\Big|_{P_i}^{P_f} = 0. \tag{30}$$

That is, there is no change in $\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi$:

$$\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi = \text{constant}, \tag{31}$$

or

$$\text{kinetic energy} + \text{``potential energy''} = \text{constant}, \tag{32}$$

and (32) explains why such a force field is called conservative because the total energy (kinetic plus potential) is conserved. Now we see the motivation for including the minus sign in (26), namely, so we end up with plus signs in (30)–(32). ▮

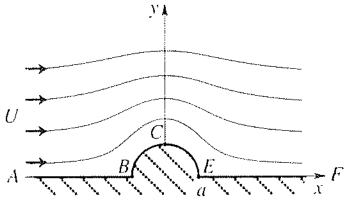


Figure 7. Flow over a building.

EXAMPLE 3. *Irrotational Fluid Flow.* Consider a building with a semicircular cross section. A cross wind, say of speed U (Fig. 7), will cause a lift force on the building (i.e., in the positive y direction) which can be quite large and which must be taken into account in the structural design. If we know the air velocity field $\mathbf{v}$ then we can use the Bernoulli equation of fluid mechanics (Exercise 12) to compute the pressure field, in particular the pressure distribution on the roof, integration of which gives the lift force. Here, we will limit our discussion to showing how to set up the field equation governing the velocity field $\mathbf{v}$.

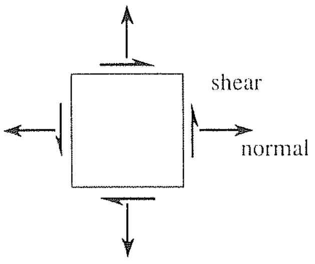


Figure 8. The stresses on a fluid particle.

Upstream (i.e., as $x \to -\infty$) the flow is simply the undisturbed free stream $U\hat{\mathbf{i}}$. Since $\nabla \times U\hat{\mathbf{i}} = 0$, we see that the flow is irrotational upstream. Since $\nabla \times \mathbf{v}$ is (Section 16.5) twice the fluid angular velocity, it follows that the fluid particles initially (i.e., upstream) have no spin; they undergo pure translation. Will they acquire angular velocity as they move downstream and pass over the semicircle? Consider the stresses acting on a typical fluid particle (Fig. 8). Just as an imbalance of horizontal or vertical normal stresses will, according to Newton's second law, cause a horizontal or vertical acceleration, respectively, an imbalance of shearing stresses results in a nonzero torque and hence an angular acceleration. But shearing stresses are possible only by virtue of the fluid property known as viscosity. We propose that air has sufficiently small viscosity for us, to a good approximation, to consider it as inviscid, that is, having no viscosity at all. With no viscosity there are no shearing stresses, hence no torque, hence no angular acceleration. Thus, having started out as irrotational (i.e., upstream) the flow will remain irrotational throughout: not only does $\nabla \times \mathbf{v} \to \mathbf{0}$ as $x \to -\infty$, $\nabla \times \mathbf{v} = \mathbf{0}$ everwhere in the field. Consequently, there exists a scalar potential Φ, called the *velocity potential* in this application, such that

$$\mathbf{v} = \nabla\Phi. \tag{33}$$

What physics has thus far been left out? We have used Newton's second law and have also made an assumption on a relevant material property, the viscosity. Also relevant is the physical law of conservation of mass, which (see Example 2 in Section 16.8) can be expressed as the "continuity equation"

$$\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma \mathbf{v}) = 0, \tag{34}$$

where σ is the fluid mass density field. If, besides assuming the fluid to be inviscid we also consider it to be incompressible, then the density σ is a constant. In that case (34) reduces to $0 + \sigma\nabla \cdot \mathbf{v} = 0$, or

$$\nabla \cdot \mathbf{v} = 0. \tag{35}$$

Finally, putting (33) into (35) gives

$$\boxed{\nabla^2\Phi = 0} \tag{36}$$

so the field equation governing Φ is the famous *Laplace equation* (36).

The remainder of this example is optional since it draws upon material contained in the optional Section 16.7.

In what coordinate system should we express (36)? The uniform free stream $U\hat{\mathbf{i}}$ is expressed most readily in terms of Cartesian coordinates, but the semicircular shape suggests using polar coordinates. When we return to this problem in a later chapter, and solve it by a method of separation of variables, we will see that we need to work in polar coordinates so let us make that choice here. We consider the region to be bounded by the radial lines $\theta = 0$ (EF in Fig. 9) and $\theta = \pi$ (AB), and the semicircles $r = a$ (BCE) and $r = R$ (AGF). Since the portion $ABCEF$ is rigid, we assume that the flow is *along* that part of the boundary, neither penetrating it nor separating from it; that is, $\hat{\mathbf{n}} \cdot \mathbf{v} = \hat{\mathbf{n}} \cdot \nabla\Phi = \partial\Phi/\partial n = 0$, where the second equality is simply the directional derivative formula (7) in Section 16.4, and $\hat{\mathbf{n}}$ is the unit outward normal to $\mathcal{D}$. On BCE and on EF we have

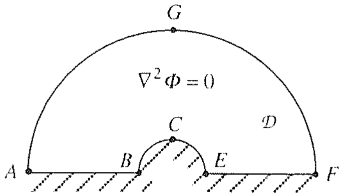


Figure 9. The domain $\mathcal{D}$.

$$\begin{aligned} BCE: &\quad \hat{\mathbf{n}} \cdot \nabla\Phi = -\hat{\mathbf{e}}_r \cdot \left(\frac{\partial\Phi}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial\Phi}{\partial\theta}\hat{\mathbf{e}}_\theta \right) = -\frac{\partial\Phi}{\partial r} = 0, \\ EF: &\quad \hat{\mathbf{n}} \cdot \nabla\Phi = -\hat{\mathbf{e}}_\theta \cdot \left(\frac{\partial\Phi}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial\Phi}{\partial\theta}\hat{\mathbf{e}}_\theta \right) = -\frac{1}{r}\frac{\partial\Phi}{\partial\theta} = 0 \end{aligned} \tag{37}$$

so $\partial\Phi/\partial r = 0$ on BCE, and $\partial\Phi/\partial\theta = 0$ on EF. Similarly, $\partial\Phi/\partial\theta = 0$ on AB.

On the semicircle AGF the boundary condition is that $\mathbf{v} \sim U\hat{\mathbf{i}}$ as $R \to \infty$. That is,

$$\nabla\Phi = \frac{\partial\Phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\Phi}{\partial y}\hat{\mathbf{j}} \sim U\hat{\mathbf{i}}$$

so $\partial\Phi/\partial x \sim U$ and $\partial\Phi/\partial y \to 0$ as $R \to \infty$.[*] Thus, $\Phi \sim Ux$, or expressing this result in terms of polar coordinates, $\Phi \sim Ur\cos\theta$ as $r \to \infty$. The resulting problem governing $\Phi(r,\theta)$ is as follows,

$$\begin{aligned}
\nabla^2\Phi &= \frac{\partial^2\Phi}{\partial r^2} + \frac{1}{r}\frac{\partial\Phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\theta^2} = 0 \qquad (a < r < \infty,\ \ 0 < \theta < \pi) \\
&\frac{\partial\Phi}{\partial\theta}(r,0) = \frac{\partial\Phi}{\partial\theta}(r,\pi) = 0 \qquad (a < r < \infty) \\
&\frac{\partial\Phi}{\partial r}(a,\theta) = 0 \qquad (0 < \theta < \pi) \\
&\Phi(r,\theta) \sim Ur\cos\theta \qquad \text{as } r \to \infty, \qquad (0 < \theta < \pi)
\end{aligned} \tag{38}$$

and is solved within Chapter 20.

COMMENT 1. More generally, any irrotational incompressible flow is governed by a Laplace equation $\nabla^2\Phi = 0$ on the velocity potential Φ, and is known as a potential flow.

COMMENT 2. How can we quantify the accuracy of our two assumptions – that the fluid is inviscid and incompressible? One learns, in a course on fluid mechanics, that the accuracy of these assumptions rests on the size of two well known nondimensional parameters: the effects of viscosity will be negligible if the *Reynolds number* is much greater than unity, and the effects of compressibility will be negligible if the square of the *Mach number* is much less than unity.

COMMENT 3. It would be natural to wonder if the velocity potential Φ admits a simple physical significance. Such significance can in fact be attached to Φ in terms of so-called impulsive pressures, discussion of which can be found in the book by Wilson.[*] ▮

16.10.2. Non-Cartesian coordinates. (Optional) Let us limit our coverage of the non-Cartesian case to one example.

EXAMPLE 4. *Cylindrical Coordinates.* Evaluate the line integral $\int_C \mathbf{v}\cdot d\mathbf{R}$, where

$$\mathbf{v} = 3r^2 z\sin\theta\hat{\mathbf{e}}_r + r^2 z\cos\theta\hat{\mathbf{e}}_\theta + (r^3\sin\theta - 3z^2)\hat{\mathbf{e}}_z \tag{39}$$

and where C is a straight line from $P_i = (x_i, y_i, z_i) = (2,0,1)$ to $P_f = (x_f, y_f, z_f) = (1,1,0)$. Recalling that in cylindrical coordinates $d\mathbf{R} = dr\,\hat{\mathbf{e}}_r + r\,d\theta\,\hat{\mathbf{e}}_\theta + z\,\hat{\mathbf{e}}_z$, we have

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \int_C 3r^2 z\sin\theta\,dr + r^3 z\cos\theta\,d\theta + (r^3\sin\theta - 3z^2)\,dz. \tag{40}$$

Using the cylindrical coordinate expression for $\nabla\times\mathbf{v}$, we find that $\nabla\times\mathbf{v} = \mathbf{0}$ so we need not evaluate the integral directly; we can use either path simplification or the potential method.

[*] $f(x) \sim g(x)$ as $x \to x_0$ means that $f(x)/g(x) \to 1$ as $x \to x_0$. We never write $f(x) \sim 0$ as $x \to x_0$ because $f(x)/0$ cannot tend to 1; it is better notation to write $f(x) \to 0$ rather than $f(x) \sim 0$. That is why we wrote $\partial\Phi/\partial y \to 0$ rather than $\partial\Phi/\partial y \sim 0$.

[*] D. H. Wilson, *Hydrodynamics* (London: Edward Arnold, 1959), Sec.10.

Since C is already a straight line, isn't it already simple? No, we can do better for the simplest path is made up of one or more segments on which only one coordinate varies. Specifically, let us use the path C_1, C_2, C_3 shown in Fig. 10. On C_1 only r varies (from 2 to $\sqrt{2}$) while θ and z are fixed ($\theta = 0$ and $z = 1$); on C_2 only θ varies (from 0 to $\pi/4$) while r and z are fixed ($r = \sqrt{2}$ and $z = 1$); and on C_3 only z varies (from 1 to 0) while r and θ are fixed ($r = \sqrt{2}$ and $\theta = \pi/4$) so

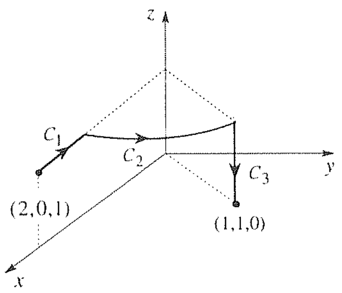


Figure 10. The modified path.

$$\begin{aligned}\int_C \mathbf{v}\cdot d\mathbf{R} &= \int_{C_1} 3r^2 z\sin\theta\, dr\Big|_{\theta=0,z=1} + \int_{C_2} r^3 z\cos\theta\, d\theta\Big|_{r=\sqrt{2},z=1} \\ &\quad + \int_{C_3} (r^3\sin\theta - 3z^2)\, dz\Big|_{r=\sqrt{2},\theta=\pi/4} \\ &= \int_2^{\sqrt{2}} 0\, dr + \int_0^{\pi/4} 2\sqrt{2}\,\cos\theta\, d\theta + \int_1^0 (2-3z^2)\, dz \\ &= 1. \end{aligned} \tag{41}$$

Alternatively, let us use the potential method. Write out the r, θ, z components of $\mathbf{v} = \nabla\Phi$:

$$\frac{\partial\Phi}{\partial r} = 3r^2 z\sin\theta, \tag{42a}$$

$$\frac{1}{r}\frac{\partial\Phi}{\partial\theta} = r^2 z\cos\theta, \tag{42b}$$

$$\frac{\partial\Phi}{\partial z} = r^3\sin\theta - 3z^2. \tag{42c}$$

Integrating (42a) gives

$$\Phi(r,\theta,z) = \int 3r^2 z\sin\theta\,\partial r = r^3 z\sin\theta + A(\theta,z), \tag{43}$$

and putting (43) into (42b) gives

$$\frac{1}{r}\left(r^3 z\cos\theta + \frac{\partial A}{\partial\theta}\right) = r^2 z\cos\theta$$

so that $\partial A/\partial\theta = 0$. Thus $A(\theta,z)$ does not vary with θ so it can be expressed as a function of z only, say $B(z)$. Then (43) becomes $\Phi(r,\theta,z) = r^3 z\sin\theta + B(z)$, and putting this expression into (42c) gives $r^3\sin\theta + B'(z) = r^3\sin\theta - 3z^2$. Thus, $B'(z) = -3z^2$ so $B(z) = -z^3 + C$, where C is an arbitrary constant. The result is that

$$\Phi(r,\theta,z) = r^3 z\sin\theta - z^3 + C \tag{44}$$

where we can, without loss, set $C = 0$, say. Finally,

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f} = (r^3 z\sin\theta - z^3)\Big|_{r=2,\theta=0,z=1}^{r=\sqrt{2},\theta=\pi/4,z=0} = 1, \tag{45}$$

in agreement with (41). ▮

EXAMPLE 5. *Role of Simple Connectedness.* Our purpose in this final example is to indicate the significance of the requirement, in Theorem 16.10.1, that the domain $\mathcal{D}$ be simply connected. Consider the vector field

$$\mathbf{H} = \frac{I}{2\pi r}\,\hat{\mathbf{e}}_\theta, \tag{46}$$

which is the magnetic field intensity induced in 3-space by a current I flowing in a wire which extends from $z = -\infty$ to $z = +\infty$ along the z axis (Fig. 11). Since $H_r = 0$, $H_z = 0$, and $H_\theta = I/2\pi r$ is a function of r only, (17) in Section 16.7 gives

$$\begin{aligned}\nabla \times \mathbf{H} &= \frac{1}{r}\frac{\partial}{\partial r}(rH_\theta)\,\hat{\mathbf{e}}_z \\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{I}{2\pi}\right)\hat{\mathbf{e}}_z = \frac{1}{r}(0)\,\hat{\mathbf{e}}_z = \mathbf{0}\end{aligned} \tag{47}$$

so the field is irrotational. Yet,

$$\begin{aligned}\oint_{C_1} \mathbf{H}\cdot d\mathbf{R} &= \oint_{C_1} \frac{I}{2\pi r}\,\hat{\mathbf{e}}_\theta \cdot (dr\,\hat{\mathbf{e}}_r + r\,d\theta\,\hat{\mathbf{e}}_\theta + dz\,\hat{\mathbf{e}}_z) \\ &= \oint_{C_1} \frac{I}{2\pi}\,d\theta = \frac{I}{2\pi}\,2\pi = I\end{aligned} \tag{48}$$

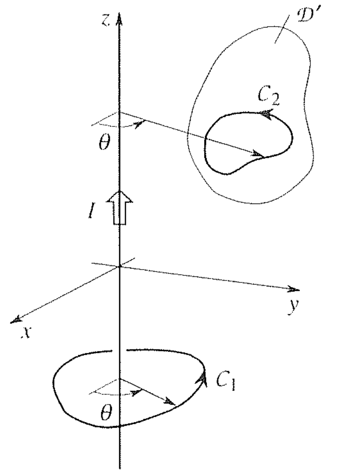


Figure 11. Current-carrying wire.

is nonzero, in apparent contradiction of Theorem 16.10.1. However, observe that C_1 does not lie within a *simply connected* domain $\mathcal{D}$ throughout which $\mathbf{H}$ is C^1 and $\nabla \times \mathbf{H} = \mathbf{0}$. That is, $\mathbf{H} = (I/2\pi r)\,\hat{\mathbf{e}}_\theta$ is not even defined along the z axis ($r = 0$), let alone C^1 there. Similarly, the last equality in (47) holds only for $r \neq 0$. Thus, $\mathbf{H}$ is C^1 and $\nabla \times \mathbf{H} = \mathbf{0}$ in a region $\mathcal{D}$ consisting of all 3-space but with the z axis excluded, that is, the annulus $0 < r < \infty$. That domain is not simply connected so the result (48) does not violate the theorem.

Understand, in (48), that $\oint_{C_1} d\theta = 2\pi$ because θ increases by 2π as C_1 is traversed: $\theta_f = \theta_i + 2\pi$. However, for any curve that does not encircle the z axis, such as C_2 (Fig. 11), $\theta_f = \theta_i$ so

$$\oint_{C_2} \mathbf{H}\cdot d\mathbf{R} = \oint_{C_2} \frac{I}{2\pi}\,d\theta = 0. \tag{49}$$

The result (49) follows from Theorem 16.10.1 because there *is* a simply connected domain $\mathcal{D}'$ (Fig. 11) containing C_2, throughout which $\mathbf{H}$ is C^1 and irrotational. That is, $\mathcal{D}'$ is not pierced by the z axis. ▮

Closure. The focal point of this section is the irrotational field theorem, Theorem 16.10.1, in which each of the four stated items implies the other three. However, in applications the direction of inference is usually as follows: if $\mathbf{v}$ is irrotational, then there exists a scalar potential Φ such that $\mathbf{v} = \nabla\Phi$, the line integral $\oint_C \mathbf{v}\cdot d\mathbf{R}$ around every closed path is zero, and the line integral $\int_C \mathbf{v}\cdot d\mathbf{R}$ along every open path C is path independent and depends only upon the endpoints of C. In particular, we show that if C is open, then the integral is simply the change in the potential between the endpoints P_i and P_f:

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f}. \tag{50}$$

Thus, if $\mathbf{v}$ is irrotational (and C^1 in a simply connected domain $\mathcal{D}$) then we can compute the line integral $\int_C \mathbf{v} \cdot d\mathbf{R}$ along an open path C in $\mathcal{D}$ either by simplifying the path or by finding the potential function Φ and using (50). To find Φ, we integrate the three scalar components of $\mathbf{v} = \nabla \Phi$.

An important advantage of working with Φ, rather than $\mathbf{v}$, is that it is a single scalar function, whereas $\mathbf{v}$ is comprised of three scalar functions, namely, its three components. That scalar potentials are of great importance is witnessed by the fact that many of them are so well known, for instance, the gravitational potential induced by a distribution of mass, the electric potential (i.e., the voltage) induced by a distribution of charge, the entropy in thermodynamics (introduced in the exercises), and the velocity potential of fluid mechanics. In terms of these potentials we have such important results as the statement of conservation of energy (32) for conservative force fields and the Laplace equation (36) governing irrotational incompressible fluid flow.

EXERCISES 16.10

1. In what domain(s) are $\mathbf{v}$ both C^1 and irrotational?

($\underline{a}$) $\mathbf{v} = x^2\hat{\mathbf{i}} - 3y\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
(b) $\mathbf{v} = 2\sin x\hat{\mathbf{i}} + y^3\hat{\mathbf{j}} + x\hat{\mathbf{k}}$
(c) $\mathbf{v} = e^x\hat{\mathbf{i}} - (2/y)\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(d) $\mathbf{v} = (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})/(x^2 + y^2 + z^2)$
($\underline{e}$) $\mathbf{v} = (-y\hat{\mathbf{i}} + x\hat{\mathbf{j}})/(x^2 + y^2)$
(f) $\mathbf{v} = x^3\hat{\mathbf{i}} - (z\hat{\mathbf{j}} - y\hat{\mathbf{k}})/(y^2 + z^2)$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + y\hat{\mathbf{j}} - x\hat{\mathbf{k}}$

2. Show whether or not $\mathbf{v}$ is irrotational; if it is, find its potential Φ such that $\mathbf{v} = \nabla\Phi$. In what domain $\mathcal{D}$ is your result valid?

($\underline{a}$) $\mathbf{v} = \hat{\mathbf{i}} - 2\hat{\mathbf{j}} - 8\hat{\mathbf{k}}$
(b) $\mathbf{v} = x^3\hat{\mathbf{i}} - y\hat{\mathbf{j}} + \sin z\hat{\mathbf{k}}$
(c) $\mathbf{v} = z\hat{\mathbf{i}} + y^3\hat{\mathbf{j}} + x\hat{\mathbf{k}}$
(d) $\mathbf{v} = x(2z + y)\hat{\mathbf{i}} + \dfrac{x^2}{2}\hat{\mathbf{j}} + x^2\hat{\mathbf{k}}$
(e) $\mathbf{v} = ye^{-x}\hat{\mathbf{i}} + e^{-|x|}\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
($\underline{f}$) $\mathbf{v} = 2ye^x\hat{\mathbf{i}} + 2e^{-|x|}\hat{\mathbf{j}} - 3\hat{\mathbf{k}}$
(g) $\mathbf{v} = 2xz\hat{\mathbf{i}} + 3y\hat{\mathbf{j}} + x^2\hat{\mathbf{k}}$
(h) $\mathbf{v} = z\hat{\mathbf{i}} + z^2\hat{\mathbf{j}} + (x + 2yz)\hat{\mathbf{k}}$
(i) $\mathbf{v} = 2ze^{2y}\hat{\mathbf{j}} + e^{2y}\hat{\mathbf{k}}$
(j) $\mathbf{v} = 2xy\hat{\mathbf{i}} + (2ze^{2y} + x^2)\hat{\mathbf{j}} + e^{2y}\hat{\mathbf{k}}$
($\underline{k}$) $\mathbf{v} = xy\hat{\mathbf{i}} - y\hat{\mathbf{k}}$
(l) $\mathbf{v} = 6\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + xyz\hat{\mathbf{k}}$
(m) $\mathbf{v} = 2xz\hat{\mathbf{i}} + x^2\hat{\mathbf{k}}$
(n) $\mathbf{v} = xz\hat{\mathbf{i}} + 2(y+1)z\hat{\mathbf{j}} + \left(\dfrac{x^2}{2} + y^2 + 2y\right)\hat{\mathbf{k}}$
(o) $\mathbf{v} = yz\hat{\mathbf{i}} + xz\hat{\mathbf{j}} + xy\hat{\mathbf{k}}$

3. Can functions $f(x, y, z)$, $g(x, y, z)$ be found such that $\mathbf{v}$ is irrotational? If so, find one such f and g. If not, why not?

($\underline{a}$) $\mathbf{v} = xy\hat{\mathbf{i}} - z\hat{\mathbf{j}} + f\hat{\mathbf{k}}$
(b) $\mathbf{v} = 2xy\hat{\mathbf{i}} + x^2\hat{\mathbf{j}} + f\hat{\mathbf{k}}$
(c) $\mathbf{v} = f\hat{\mathbf{i}} + 2xyz\hat{\mathbf{j}} + g\hat{\mathbf{k}}$
(d) $\mathbf{v} = z^2\hat{\mathbf{i}} + f\hat{\mathbf{j}} + g\hat{\mathbf{k}}$

4. Assuming that the conditions of Theorem 16.10.1 are met, are the following correct? Explain.

(a) $$\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_i, z_i)\,dx + \int_{y_i}^{y_f} v_y(x_f, y, z_i)\,dy + \int_{z_i}^{z_f} v_z(x_f, y_f, z)\,dz$$

(b) $$\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_f, z_f)\,dx + \int_{y_i}^{y_f} v_y(x_i, y, z_i)\,dy + \int_{z_i}^{z_f} v_z(x_i, y_f, z)\,dz$$

(c) $$\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_i, z_i)\,dx + \int_{y_i}^{y_f} v_y(x_f, y, z_f)\,dy + \int_{z_i}^{z_f} v_z(x_i, y_f, z)\,dz$$

5. Consider $\int_C \mathbf{v}\cdot d\mathbf{R}$, where C is the path $x = \sin\tau$, $y = \cos 3\tau$, $z = 2\tau$ from $\tau = 0$ to $\tau = \pi$. Show that $\mathbf{v}$ is irrotational, and use this fact to evaluate the integral two ways:

by path simplification and by the potential method. Sketch the simplified path.

(a) $\mathbf{v} = y\hat{\mathbf{i}} + x\hat{\mathbf{j}}$
(b) $\mathbf{v} = 3x^2 \cos 2y\hat{\mathbf{i}} - 2x^3 \sin 2y\hat{\mathbf{j}}$
(c) $\mathbf{v} = 5x^{3/2}e^{2y}\hat{\mathbf{i}} + (4x^{5/2}e^{2y} + 5y^2)\hat{\mathbf{j}}$
(d) $\mathbf{v} = 3\hat{\mathbf{i}} + 6y^2z^{5/2}\hat{\mathbf{j}} + (5y^3z^{3/2} + 2)\hat{\mathbf{k}}$
(e) $\mathbf{v} = (y^3 - \cos z)\hat{\mathbf{j}} + (y \sin z + 2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = 2x^2\hat{\mathbf{i}} - 2yz\hat{\mathbf{j}} - (y^2 + 3)\hat{\mathbf{k}}$

6. Repeat Exercise 5, but where C is the path $x = \cos^3 \tau$, $y = \tau^2$, $z = 3\tau$ from $\tau = 0$ to $\tau = \pi$.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(b) $\mathbf{v} = 8\hat{\mathbf{i}} + (z^2 - 3y)\hat{\mathbf{j}} + (2zy - \sqrt{z})\hat{\mathbf{k}}$
(c) $\mathbf{v} = (x^2 - 2y^{3/2})\hat{\mathbf{i}} - 3x\sqrt{y}\hat{\mathbf{j}} + e^z\hat{\mathbf{k}}$
(d) $\mathbf{v} = y^2z^2\hat{\mathbf{i}} + 2xyz^2\hat{\mathbf{j}} + 2xy^2z\hat{\mathbf{k}}$

7. (*Exact differentials*) An expression $P\,dx + Q\,dy + R\,dz$, where P, Q, and R are functions of x, y, z defined in some domain $\mathcal{D}$, is called a *first-order differential form* in three variables. If there exists a function f such that $P\,dx + Q\,dy + R\,dz = df$ in $\mathcal{D}$, then the form is said to be an **exact differential**. Assuming that P, Q, R are C^1 in $\mathcal{D}$, show that for $P\,dx + Q\,dy + R\,dz$ to be an exact differential it is necessary and sufficient that $\nabla \times (P\hat{\mathbf{i}} + Q\hat{\mathbf{j}} + R\hat{\mathbf{k}}) = \mathbf{0}$ in $\mathcal{D}$. Further, show that f is the scalar potential of $P\hat{\mathbf{i}} + Q\hat{\mathbf{j}} + R\hat{\mathbf{k}}$.

8. (*Stream function*) Consider the plane flow $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$ of an incompressible fluid, where $\mathbf{v}$ is C^1. The volume flow rate (in meters^3/second, say) crossing the curve OP from left to right is

$$Q = \int_C v_x\,dy - v_y\,dx \equiv \int_C \mathbf{f} \cdot d\mathbf{R}, \tag{8.1}$$

where $\mathbf{f} = -v_y(x, y)\hat{\mathbf{i}} + v_x(x, y)\hat{\mathbf{j}}$. Since $\nabla \times \mathbf{f} = (\partial v_x/\partial x + \partial v_y/\partial y)\hat{\mathbf{k}}$ is zero by virtue of the continuity equation [(29) in Section 16.8, with $v_z = 0$ in this case], it follows that

$$Q = \int_{(0,0)}^{(x,y)} v_x\,dy - v_y\,dx \tag{8.2}$$

is a (single-valued) function of x and y, say $\Psi(x, y)$, where

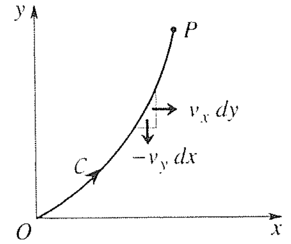
$$\frac{\partial \Psi}{\partial x} = -v_y, \qquad \frac{\partial \Psi}{\partial y} = v_x. \tag{8.3}$$

Ψ is called the **stream function**.

(a) Show that the $\Psi(x, y) =$ constant curves are the streamlines, as were defined in Example 2 of Section 16.2.

(b) If $\mathbf{v} = 2y\hat{\mathbf{i}}$, verify that $\nabla \cdot \mathbf{v} = 0$ (so that Ψ exists), and determine $\Psi(x, y)$. NOTE: Like the velocity potential Φ, the stream function Ψ can be determined only to within an arbitrary additive constant.

(c) Repeat part (b), for $\mathbf{v} = 2xe^{2y}\hat{\mathbf{i}} - e^{2y}\hat{\mathbf{j}}$. Sketch the streamline through $(1, 0)$.



(d) If $\nabla \times \mathbf{v} \equiv \Omega(x, y)\hat{\mathbf{k}}$, show that

$$\nabla^2\Psi = -\Omega(x, y). \tag{8.4}$$

NOTE: If the flow is incompressible *and* irrotational (so that $\Omega = 0$), then there exist both a velocity potential Φ satisfying $\nabla^2\Phi = 0$, *and* a streamfunction Ψ satisfying $\nabla^2\Psi = 0$. If the flow is incompressible but *not* irrotational, Φ does not exist – yet the stream function does exist and satisfies (8.4).

(e) If the flow is incompressible and irrotational, so that both Ψ and Φ exist, show that the $\Psi =$ constant curves and the $\Phi =$ constant curves are everywhere *orthogonal*; i.e., at each point $P = (x, y)$ the $\Psi =$ constant curve through P, and the $\Phi =$ constant curve through P intersect at a right angle. HINT: Consider $\nabla\Psi \cdot \nabla\Phi$.

9. (*Entropy of an ideal gas*) Any gas satisfying the equation of state $pv = RT$, where p is the pressure, v is the volume per mole, T is the absolute temperature, and R is the universal gas constant, is said to be an *ideal gas*. The first law of thermodynamics for one mole of an ideal gas can be expressed as

$$dq = p\,dv + c_v\,dT, \tag{9.1}$$

where dq is the heat input and $c_v = c_v(T)$ is the specific heat at constant volume. Show that the right-hand side of (9.1) is *not* an exact differential (see Exercise 7), but that if we multiply through by $1/T$, then

$$\frac{dq}{T} = \frac{p}{T}\,dv + \frac{c_v}{T}\,dT \tag{9.2}$$

is an exact differential, say $dq/T \equiv ds$, where $s(v, T)$ is called the *entropy* of the (one mole of) gas. [Since ds is an exact differential, s is indeed a *function* of v, T. That is, it is uniquely determined, for the given gas, by the point (v, T)

in the v, T plane (i.e., by the state), independent of the path history.]

10. (*Solenoidal fields*) Let $\mathbf{v}$ be a vector field defined in a domain $\mathcal{D}$. If $\nabla \cdot \mathbf{v} = 0$ at each point in $\mathcal{D}$, then $\mathbf{v}$ is said to be **solenoidal** in $\mathcal{D}$. We have the following companion to Theorem 16.10.1:

THEOREM 16.10.2 *Solenoidal Field*
If $\mathbf{v}$ is C^1 and solenoidal in a simply connected domain $\mathcal{D}$, then there exists a vector field $\mathbf{w}$ in $\mathcal{D}$ such that $\mathbf{v} = \nabla \times \mathbf{w}$.

Prove this theorem, for the case where $\mathcal{D}$ is the prism $x_1 < x < x_2$, $y_1 < y < y_2$, $z_1 < z < z_2$. HINT: We need to show that the equation $\mathbf{v} = \nabla \times \mathbf{w}$ does admit a solution $\mathbf{w}$. To do this, start with the three scalar equations

$$\frac{\partial w_z}{\partial y} - \frac{\partial w_y}{\partial z} = v_x, \qquad \frac{\partial w_x}{\partial z} - \frac{\partial w_z}{\partial x} = v_y, \qquad \frac{\partial w_y}{\partial x} - \frac{\partial w_x}{\partial y} = v_z. \tag{10.1}$$

There is enough leeway in (10.1) to set one of the components of $\mathbf{w}$, say w_x, equal to zero. Then the latter two equations in (10.1) become

$$\frac{\partial w_z}{\partial x} = -v_y \qquad \text{and} \qquad \frac{\partial w_y}{\partial x} = v_z, \tag{10.2}$$

which can be integrated to give w_y to within an arbitrary additive function $A(y, z)$ and w_z to within an arbitrary additive function $B(y, z)$. Putting these results into the first equation in (10.1), show that it is possible to choose $A(y, z) = 0$ so that

$$\mathbf{w} = 0\hat{\mathbf{i}} + \left[\int_{x_0}^{x} v_z(\xi, y, z)\,\partial\xi\right]\hat{\mathbf{j}} + \left[\int_{y_0}^{y} v_x(x_0, \eta, z)\,\partial\eta - \int_{x_0}^{x} v_y(\xi, y, z)\,\partial\xi\right]\hat{\mathbf{k}}, \tag{10.3}$$

where x_0, y_0 are any constants such that $x_1 \le x_0 \le x_2$ and $y_1 \le y_0 \le y_2$. Show that to the right-hand side of (10.3) we can add ∇f, the gradient of an arbitrary scalar function f that is C^2 in $\mathcal{D}$. NOTE: Observe the pattern that emerges in this section: If $\mathbf{v}$ is irrotational ($\nabla \times \mathbf{v} = \mathbf{0}$), $\mathbf{v}$ is expressible as the gradient of a scalar potential Φ ($\mathbf{v} = \nabla\Phi$). Analogously, if $\mathbf{v}$ is solenoidal ($\nabla \cdot \mathbf{v} = 0$), $\mathbf{v}$ is expressible as the curl of a vector potential $\mathbf{w}$ ($\mathbf{v} = \nabla \times \mathbf{w}$). The notions of irrotational and solenoidal fields are complementary in the sense that *every C^1 vector field $\mathbf{v}$, defined in a bounded simply connected domain $\mathcal{D}$, can be split as $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1$ is irrotational in $\mathcal{D}$ and $\mathbf{v}_2$ is solenoidal in $\mathcal{D}$*, a fact that we state without proof.

11. Show that $\mathbf{v}$ is solenoidal (see Exercise 10), and determine a vector potential $\mathbf{w}$. (As noted in Exercise 10, $\mathbf{w}$ is not uniquely determined.)

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ (a, b, c constants)
(b) $\mathbf{v} = y\hat{\mathbf{i}} + x\hat{\mathbf{j}}$
(c) $\mathbf{v} = 3\hat{\mathbf{i}} + x^2\hat{\mathbf{j}} + 4y\hat{\mathbf{k}}$
(d) $\mathbf{v} = xy\hat{\mathbf{i}} - yz\hat{\mathbf{k}}$
(e) $\mathbf{v} = 2y\hat{\mathbf{i}} + x^2y^2\hat{\mathbf{j}} - 2x^2yz\hat{\mathbf{k}}$
(f) $\mathbf{v} = x^3y\hat{\mathbf{i}} - z\hat{\mathbf{j}} - 3x^2y(z+1)\hat{\mathbf{k}}$
(g) $\mathbf{v} = x\sin y\hat{\mathbf{i}} + (\cos y + 2z)\hat{\mathbf{j}} + x^2y\hat{\mathbf{k}}$

12. (*Integrating equations of motion*) In Example 3 we considered a single particle. Here we consider a continuum of particles. Consider the motion of a fluid, assumed irrotational and incompressible, in a domain $\mathcal{D}$, under the action of a uniform gravitational force field $-g\hat{\mathbf{k}}$ (per unit mass). If σ is the mass density, $\mathbf{v}$ is the velocity field, and p is the pressure field, then we state, without derivation, that Newton's second law leads to the *Euler equation*

$$\sigma\frac{d\mathbf{v}}{dt} = -\nabla p - \sigma g\hat{\mathbf{k}}, \tag{12.1}$$

or

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} + \nabla\left(\frac{p}{\sigma}\right) + g\hat{\mathbf{k}} = \mathbf{0}, \tag{12.2}$$

where the d/dt in (12.1) is the convective derivative defined in Exercise 5 of Section 16.4. Use equation (8) in Section 16.6 to show that $(\mathbf{v}\cdot\nabla)\mathbf{v} = \nabla(\frac{1}{2}v^2)$, where v is the speed, i.e., the norm of the velocity $\mathbf{v}$, and show that it follows from (12.2) that

$$\frac{\partial\Phi}{\partial t} + \frac{1}{2}v^2 + \frac{p}{\sigma} + gz = \text{"constant"} = F(t) \tag{12.3}$$

everywhere in $\mathcal{D}$ for arbitrary $F(t)$. Equation (12.3) is one form of the well known *Bernoulli equation*.

13. Given $\int_C \mathbf{v}\cdot d\mathbf{R}$, where C is the open path

$$r = \tau, \qquad \theta = \pi\tau/2, \qquad z = \sin\frac{\pi(\tau-1)}{2}$$

for $\tau : 1 \to 2$, evaluate I both by path simplification and by the potential method, and justify the use of those methods.

(a) $\mathbf{v} = z\hat{\mathbf{e}}_r + r\hat{\mathbf{e}}_z$
(b) $\mathbf{v} = z\sin\theta\hat{\mathbf{e}}_r + z\cos\theta\hat{\mathbf{e}}_\theta + r\sin\theta\hat{\mathbf{e}}_z$
(c) $\mathbf{v} = \cos 5\theta\hat{\mathbf{e}}_r - 5\sin 5\theta\hat{\mathbf{e}}_\theta + z^2\hat{\mathbf{e}}_z$
(d) $\mathbf{v} = 2(rz - \cos\theta)\hat{\mathbf{e}}_r + 2\sin\theta\hat{\mathbf{e}}_\theta + r^2\hat{\mathbf{e}}_z$
(e) $\mathbf{v} = \rho^5\hat{\mathbf{e}}_\rho$
(f) $\mathbf{v} = \sin\phi\hat{\mathbf{e}}_\rho + \cos\phi\hat{\mathbf{e}}_\phi$
(g) $\mathbf{v} = 3\rho^2\cos\theta\hat{\mathbf{e}}_\rho - \rho^2\dfrac{\cos\theta}{\sin\phi}\hat{\mathbf{e}}_\theta$
(h) $\mathbf{v} = \dfrac{\cos\phi}{\rho}\hat{\mathbf{e}}_\phi - \dfrac{\sin\theta}{\rho\sin\phi}\hat{\mathbf{e}}_\theta$

14. In (9), we showed that

$$\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \int_{P_i}^{P_f} \nabla\Phi\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f}$$

by expressing $\nabla\Phi$ and $d\mathbf{R}$ in Cartesian coordinates. Show that we obtain the same result, namely $\Phi|_{P_i}^{P_f}$, if we use cylindrical coordinates or spherical coordinates.

Chapter 16 Review

We begin by defining the divergence of a vector field $\mathbf{v}$ as

$$\operatorname{div}\mathbf{v} = \lim_{B\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}.$$

Carrying out that limit in Cartesian coordinates gives

$$\operatorname{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z},$$

which we identify as $\nabla\cdot\mathbf{v}$, where ∇ is the vector differential operator

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}.$$

With ∇ in hand, we use it to define

$$\operatorname{grad} u = \nabla u \qquad \text{and} \qquad \operatorname{curl}\mathbf{v} = \nabla\times\mathbf{v}.$$

Thinking of $\mathbf{v}$ as a fluid velocity field (even if it is not), we can interpret $\operatorname{div}\mathbf{v}$ and $\operatorname{curl}\mathbf{v}$ as the outflow per unit volume and twice the fluid angular velocity, respectively. With the help of the directional derivative formula,

$$\frac{du}{ds} = \nabla u\cdot\hat{\mathbf{s}},$$

we learn that ∇u at any point P is normal to the u = constant surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.

Studying combinations (i.e., more than one operator and/or more than one field), we develop a number of identities such as

$$\nabla\cdot(u\mathbf{v}) = \nabla u\cdot\mathbf{v} + u\nabla\cdot\mathbf{v},$$

and meet the Laplace operator

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

which is known as the Laplacian. In particular, we find that the divergence of every curl is zero (mnemonically, "dc"), and the curl of every gradient is zero ("cg"):

$$\begin{aligned} \text{div curl } \mathbf{v} &= \nabla \cdot \nabla \times \mathbf{v} = 0, \\ \text{curl grad } u &= \nabla \times \nabla u = \mathbf{0}. \end{aligned}$$

In the optional Section 16.7 we derive formulas for the gradient, divergence, curl, and Laplacian in cylindrical and spherical coordinates. Three crucial points, in that section, are as follows. First, the partial derivatives in the ∇, within $\nabla \cdot \mathbf{v}$ say, act not only on the scalar components of $\mathbf{v}$ but also on the base vectors in $\mathbf{v}$. Second, in terms such as

$$\hat{\mathbf{e}}_r \frac{\partial}{\partial r} \cdot (v_\theta \hat{\mathbf{e}}_\theta)$$

we need to do the derivative (of $v_\theta \hat{\mathbf{e}}_\theta$) before the dot or cross product. Third, whereas

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= (u_1 \hat{\mathbf{e}}_1 + u_2 \hat{\mathbf{e}}_2 + u_3 \hat{\mathbf{e}}_3) \times (v_1 \hat{\mathbf{e}}_1 + v_2 \hat{\mathbf{e}}_2 + v_3 \hat{\mathbf{e}}_3) \\ &= \begin{vmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \end{aligned}$$

holds for any vectors $\mathbf{u}$ and $\mathbf{v}$ (where $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ are the right-handed orthonormal base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, or $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ or $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$), it does *not* hold, for cylindrical or spherical coordinates, if $\mathbf{u}$ is the ∇ operator.

In Sections 16.8 and 16.9 we study the Gauss divergence theorem and Stokes's theorem (as well as Green's theorem and Green's identities) and observe the pattern of being able to express an integral over an n-dimensional integral as an integral (or evaluation) over its $(n-1)$-dimensional boundary, provided that the integrand is a sort of derivative:

$$\begin{aligned} 1 \to 0: &\quad \int_a^b F'(x)\, dx = F(x)\Big|_{x=a}^{x=b} &&\text{(Fundamental Theorem)} \\ 2 \to 1: &\quad \int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \oint_C \mathbf{v} \cdot d\mathbf{R} &&\text{(Stokes's Theorem)} \\ 3 \to 2: &\quad \int_V \nabla \cdot \mathbf{v}\, dV = \int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &&\text{(Divergence Theorem)} \end{aligned}$$

where the numbers at the left denote the dimensions n and $n-1$. [In fact, the Fundamental Theorem is just a one-dimensional version of the Divergence Theorem, where $\mathbf{v} = F(x)\hat{\mathbf{i}}$ and where the region V is a rectangular prism extending from $x = a$ to $x = b$ in the x-direction, for then $\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV$ reduces to $\int_a^b F'(x)\, dx$ times the area A of the end faces at $x = a$ and $x = b$, and $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ reduces to $F(b) - F(a)$ times A; then A cancels from the left- and right-hand sides.] These theorems are especially useful in deriving various field equations – such as the continuity equation of fluid mechanics, the diffusion equation, and various Maxwell's equations.

In the final section, on irrotational fields, the key result is Theorem 16.10.1, which (under the conditions stated therein) expressed the equivalance between $\mathbf{v}$ being irrotational, the existence of a scalar potential Φ, the line integral around every closed loop being zero, and the line integral on an open path being independent of the path.

Chapter 17

Fourier Series, Fourier Integral, Fourier Transform

17.1 Introduction

This chapter is about a number of methods associated with the French applied mathematician *Joseph Fourier* (1768–1830), methods which will be central when we turn next to the subject of partial differential equations in Chapters 18–20. In fact, series of trigonometric functions, which we now call **Fourier series**, were already being actively studied – by Euler, Lagrange, d'Alembert, Daniel Bernoulli, and others – even when Fourier was born. For instance, d'Alembert had already given integral formulas for computing the coefficients in those series, and various specific "Fourier series" had been put forward. Yet, there was considerable debate regarding the class of functions that could be successfully expanded in such series, and various sets of sufficient conditions were slow to appear, the first being given by Peter Gustav Lejeune–Dirichlet around 1829. Part of the difficulty was that even the meaning of the term "function" was not yet clear or agreed upon. But it is also true that mathematical issues were deep and elusive. For instance, in 1873, Paul Du Bois–Reymond put forward an example of a function that is continuous on $(-\pi, \pi)$, yet has a Fourier series that fails to converge at any point in that interval! Thus, the subject of Fourier series has been one of the most fertile in the development of modern pure and applied mathematics.

Although Fourier neither invented "Fourier" series nor settled the outstanding fundamental questions, he did use them fruitfully, especially for problems regarding the conduction of heat in solids, governed by the PDE

$$\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t}$$

that is derived in Section 16.8 and to which we return in Chapter 18. In claiming that an "arbitrary" function could be represented by a Fourier series Fourier overstated the case, and his work was faulted for lack of rigor. Yet he had the insight to see the power of these new methods. His work advanced the use of Fourier methods and solidified techniques that would be needed to solve problems in field theory and that continue to be developed even today.

To explain what this chapter is about we ask you to recall, first, the importance of being able to expand a given vector in terms of a set of orthogonal base vectors. Likewise, we will find in this chapter that a given (sufficiently well-behaved) function can be expanded in terms of a set of "orthogonal functions." Fourier and his contemporaries did not have vector space concepts available to them but, essentially, they were seeking sets of orthogonal functions to be used as base vectors in an infinite-dimensional function space. That task is much more difficult than for finite-dimensional spaces. For instance, if we have n orthogonal vectors in an n-dimensional space (where n is finite), then they provide a basis. However, in an infinite-dimensional space an infinite set of orthogonal functiôns may, but need not, be a basis. For suppose we have an infinite set of orthogonal functions that *is* a basis. If we remove one of them, then we have an infinite number of them left, but that diminished set will not be a basis. Further, expansions in an infinite-dimensional space are infinite series, so subtle matters arise regarding their convergence and manipulation.

Fourier series are introduced and developed in Section 17.3, though not from a vector space point of view. The alternative, and more modern, vector space approach is given subsequently in Section 17.6. By then, the nagging question arises: Where does the set of orthogonal functions, which comprise the individual terms in a Fourier series, "come from"? Are there other such sets? The answer is provided, in Sections 17.7 and 17.8, by the Sturm–Liouville theory. There, it is revealed that such sets of orthogonal base vectors are generated as the eigenfunctions of second-order differential equation eigenvalue problems known as Sturm–Liouville problems. Thus, Sections 17.7 and 17.8 are the function space analog of Section 11.3 on symmetric matrices, wherein we found that the eigenvectors of a real symmetric $n \times n$ matrix provide an orthogonal basis for n-space.

In Section 17.9 we will let the period of the periodic functions under consideration tend to infinity, and will find that the Fourier series representation gives way to a Fourier integral representation. The Fourier integral gives us, by a mere rearrangement, the Fourier transform, in Section 17.10, which is very much analogous to the Laplace transform that we studied in Chapter 5. In fact, in the final section, 17.11, we show how to derive the Laplace transform from the Fourier transform.

For interesting historical (and mathematical) accounts, we suggest the little book by R. L. Jeffery [*Trigonometric Series* (Toronto: University of Toronto Press, (1956)] as well as the historical treatise by Morris Kline [*Mathematical Thought from Ancient to Modern Times* (New York: Oxford, 1972)].*

*Other standard sources include H. S. Carslaw, *Introduction to the Theory of Fourier's Series and Integrals*, 3rd ed. (New York: Dover, 1930), and I. Grattan–Guinness, *Joseph Fourier, 1768–1830* (Cambridge, MA: MIT Press, 1972).

17.2 Even, Odd, and Periodic Functions

Before taking up our study of Fourier series, in the next section, we need to define even, odd, and periodic functions.

Let f be defined on an x interval, finite or infinite, that is centered at $x = 0$. We say that f is an **even** function if

$$\boxed{f(-x) = f(x),} \tag{1}$$

and an **odd** function if

$$\boxed{f(-x) = -f(x),} \tag{2}$$

for all x in that interval. That is, the graph of f is *symmetric* about $x = 0$ if f is even, and *antisymmetric* about $x = 0$ if f is odd. Examples are shown in Fig. 1. For example, $5, x^2, 3x^4, \cos x, \sin|x|$, and e^{-x^2} are even, and $x, 3x^3, 2x^5$, $\sin x$, and $x \cos x$ are odd.

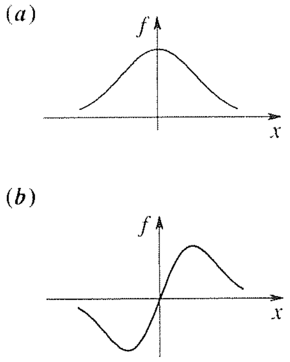


Figure 1. Even and odd.

There are several useful algebraic properties of even and odd functions, such as the following:

$$\text{even} + \text{even} = \text{even}, \tag{3a}$$
$$\text{even} \times \text{even} = \text{even}, \tag{3b}$$
$$\text{odd} + \text{odd} = \text{odd}, \tag{3c}$$
$$\text{odd} \times \text{odd} = \text{even}, \tag{3d}$$
$$\text{even} \times \text{odd} = \text{odd}. \tag{3e}$$

To prove (3e), for example, let $F(x)$ be even and let $G(x)$ be odd. Then $F(-x)G(-x) = F(x)[-G(x)] = -F(x)G(x)$, in accord with (2).

In addition, two useful integral properties are as follows. If f is even, then

$$\boxed{\int_{-A}^{A} f(x)\,dx = 2\int_{0}^{A} f(x)\,dx, \qquad (f \text{ even})} \tag{4a}$$

and if f is odd, then

$$\boxed{\int_{-A}^{A} f(x)\,dx = 0, \qquad (f \text{ odd})} \tag{4b}$$

for if we interpret the integrals in (4a) as areas (positive above the x axis, negative below it), then the area $\int_{-A}^{0} f(x)\,dx$ is equal to the $\int_{0}^{A} f(x)\,dx$ due to the symmetry of the graph of f. And in the case of (4b) the areas $\int_{-A}^{0} f(x)\,dx$ and $\int_{0}^{A} f(x)\,dx$ are negatives of each other, due to the antisymmetry of the graph of f, and hence cancel.

Alternatively, (4a) and (4b) follow directly from (1) and (2), respectively. For example, if f is odd, then

$$\begin{aligned}\int_{-A}^{A} f(x)\,dx &= \int_{-A}^{0} f(x)\,dx + \int_{0}^{A} f(x)\,dx \\ &= \int_{A}^{0} f(-t)\,(-dt) + \int_{0}^{A} f(x)\,dx \qquad (x=-t) \\ &= \int_{0}^{A} f(-t)\,dt + \int_{0}^{A} f(x)\,dx \\ &= \int_{0}^{A} -f(t)\,dt + \int_{0}^{A} f(x)\,dx \qquad (\text{oddness of } f) \\ &= -\int_{0}^{A} f(x)\,dx + \int_{0}^{A} f(x)\,dx \qquad (t=x) \\ &= 0, \end{aligned} \tag{5}$$

as stated in (4b).

Note carefully that a given function is not necessarily even or odd; it may be *both* even and odd, or it may be *neither*. Every function can be uniquely decomposed into the sum of an even function, say f_e, and an odd function, say f_o, as demonstrated by the simple identity

$$\boxed{\begin{aligned} f(x) &= \frac{f(x)+f(-x)}{2} + \frac{f(x)-f(-x)}{2} \\ &\equiv \quad f_e(x) \quad + \quad f_o(x), \end{aligned}} \tag{6}$$

for observe that

$$f_o(-x) = \frac{f(-x)-f(x)}{2} = -\frac{f(x)-f(-x)}{2} = -f_o(x)$$

and, similarly, that $f_e(-x) = f_e(x)$.

EXAMPLE 1. Surely $f(x) = e^x$ is neither even nor odd, since (Fig. 2) it is neither symmetric nor antisymmetric about $x = 0$. Putting $f(x) = e^x$ and $f(-x) = e^{-x}$ into (6) gives

$$f_e(x) = \frac{e^x + e^{-x}}{2} \qquad \text{and} \qquad f_o(x) = \frac{e^x - e^{-x}}{2}$$

as the even and odd parts of e^x, respectively. In fact, we recognize these functions as $\cosh x$ and $\sinh x$, so it is interesting that we can think of $\cosh x$ and $\sinh x$ as the even and odd parts of e^x, respectively. ▮

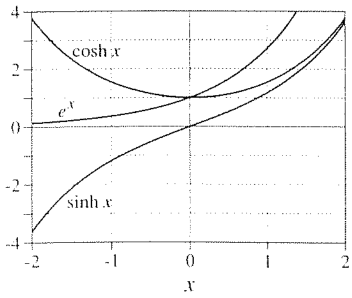


Figure 2. Even and odd parts of e^x.

Notice that (6) is reminiscent of other decompositions that occur in mathematics, whereby a mathematical object (such as a function, matrix, or vector) is broken into the sum of two complementary parts. For instance, every square matrix can be expressed as the sum of a symmetric matrix and a skew-symmetric matrix (Exercise 6, Section 10.3), every vector in 3-space can be expressed as the sum of a vector along a given line and a vector perpendicular to that line, and every C^1 vector field can be expressed as the sum of an irrotational field and a solenoidal field (Exercise 10, Section 16.10).

Next, suppose that for a given function f there exists a positive constant T such that

$$\boxed{f(x+T) = f(x)} \tag{7}$$

for every x in the domain of f. Then we say that f is a **periodic** function of x, with period T. Sometimes we say that f is **T-periodic**.

EXAMPLE 2. For example, $\sin x$ is periodic with period 2π because $\sin(x + 2\pi) = \sin x \cos 2\pi + \sin 2\pi \cos x = \sin x$ for all x. ▮

In graphical terms, one can think of the graph of a periodic function f as generated by stamping it out one period at a time, as with an inked woodblock.

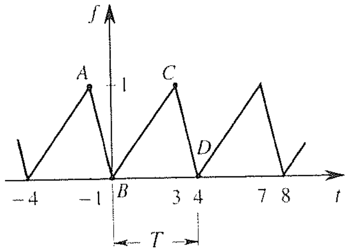


Figure 3. Periodic function f.

EXAMPLE 3. The function f shown in Fig. 3 is seen to be periodic with period $T = 4$, for if the segment BCD, for instance, is "stamped out" indefinitely to the right and left we generate the graph of f. There is nothing special about choosing the segment BCD for this purpose; ABC, or any other segment of length 4, would do as well. ▮

Notice that if f is periodic with period T, it is necessarily periodic with period $2T, 3T, 4T, \ldots$ as well. For example, $f(x+2T) = f((x+T)+T) = f(x+T) = f(x)$, so that f is periodic with period $2T$. Of all these possible periods, if there exists a *smallest* one, that period is called the **fundamental period**. Thus, $\sin x$ (in Example 2) is periodic with period $2\pi, 4\pi, 6\pi, \ldots$, and its fundamental period is 2π; f in Example 3 is periodic with period $4, 8, 12, \ldots$, and its fundamental period is 4.

In contrast, observe that if $f(x) =$ constant, then f is periodic and *every* $T > 0$ is a period. Thus, there exists no smallest period, so f does not have a fundamental period.

Closure. Even, odd, and periodic functions will be basic to our study of Fourier series, to follow. The defining properties are (1), (2), and (7), respectively.

EXERCISES 17.2

1. (a) Prove (3a). (b) Prove (3b).
(c) Prove (3c). (d) Prove (3d).

2. Provide a proof of (4a) that is analogous to the proof of (4b) given in (5).

3. Prove that

(a) f is both even and odd if and only if it is identically zero
(b) if f is even (and integrable), then $F(x) = \int_0^x f(t)\,dt$ is odd
(c) if f is odd (and integrable), then $F(x) = \int_0^x f(t)\,dt$ is even
(d) if f is even (and differentiable), then $F(x) = df/dx$ is odd
(e) if f is odd (and differentiable), then $F(x) = df/dx$ is even

4. Prove the decomposition formula (6). HINT: Start with $f(x) = f_e(x) + f_o(x)$ and change x to $-x$.

5. Determine $f_e(x)$ and $f_o(x)$. Is f even? Odd? Neither?

(a) $2 - 5x$ (b) $\sin(x + 2)$
(c) $x/(x^2 + x + 3)$ (d) xe^{-x}
(e) $x/(x + 2)$ (f) $x^2 \cos(x^3) - 8$
(g) $x^4 + x^3 + x^2 + x + 1$ (h) $\ln(1 + x^2)$
(i) $e^{-2\sin x}$ (j) $\sin(\sin x)$
(k) $\cos(\sin x)$ (l) $e^{-x}/(x^2 + 1)$

6. If F is even and G is odd, show that
(a) $1/F(x)$ is even (b) $1/G(x)$ is odd
(c) $F(G(x))$ is even (d) $G(F(x))$ is even

7. Show that if f is odd, then it is necessarily true that $f(0) = 0$.

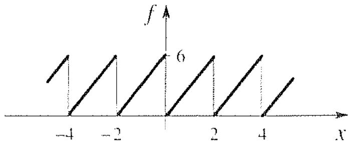
8. Let f be even (and not identically zero), and let g be odd (and not identically zero). If f and g are defined on a common interval, show that f and g are necessarily linearly independent on that interval.

9. Show that if $f(x) = g(x)$ (over a common x interval), we can equate even and odd parts: $f_e(x) = g_e(x)$, and $f_o(x) = g_o(x)$.

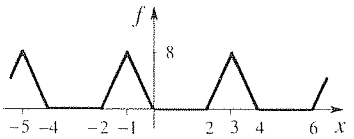
10. Show that if f is even and g is odd, and $f + g = 0$ (over a common x interval), then $f(x) = 0$ and $g(x) = 0$.

11. Determine the fundamental period in each case. Also, draw the graphs of f_e and f_o.

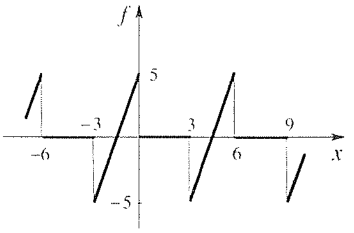
(*a*)



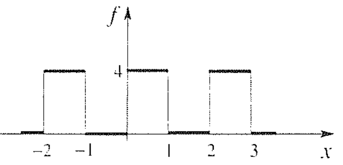
(*b*)



(*c*)



(*d*)



12. Determine whether or not the given function (defined on $-\infty < x < \infty$) is periodic. If it is, find its fundamental period (if it has one).

(a) x^4 (b) e^x (c) e^{-x}
(d) $\sin(\omega x + \phi)$ (e) $\cos 6x$ (f) $\sin 2x$
(g) $\tan x$ (h) $\sinh x$ (i) $\cosh x$
(j) $\cos^2 x$ (k) $\sin^2 x$ (l) $\sin x \cos 2x$
(m) $e^{\sin 3x}$ (n) $\sinh(\cos 2x)$ (o) $\sin(\sin x)$
(p) $\cos(\sin x)$ (q) $\sin(8\pi \cos x)$ (r) $\sin(\sin 4x)$
(s) $\sin|x|$ (t) $\cos|x|$ (u) $x \sin x$

13. The following functions are periodic. Determine the fundamental period in each case. (The a_n's and b_n's are nonzero constants.)

(a) $a_0 + a_1 \cos x$
(b) $a_0 + a_1 \cos x + a_2 \cos 2x$
(c) $6 \cos x - 4 \sin 3x$
(d) $\cos 5x + \sin 5x$
(e) $a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x$
(f) $a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$
(g) $a_0 + \sum_{n=1}^{\infty}\left(a_n \cos \dfrac{n\pi x}{\ell} + b_n \sin \dfrac{n\pi x}{\ell}\right)$

14. Show that if f is periodic with period T, then

$$\int_0^T f(x)\,dx = \int_A^{A+T} f(x)\,dx$$

for any finite value of A.

15. Show that if $f(x)$ is periodic with period T, then

(a) so is its derivative $f'(x)$

(b) so is its integral $\int_0^x f(t)\,dt$, if and only if $\int_0^T f(t)\,dt = 0$

16. Let $f(x) = 1$ whenever x is rational, and let $f(x) = 0$ whenever x is irrational. Show that f is periodic with period T where T is any (positive) rational number. (Thus there exists no smallest period, so f does not have a fundamental period.) Is it also true that f is periodic with period T, where T is any (positive) *ir*rational number? Explain.

17. Show that if f is periodic with period T, then $g(f(x))$ is, too. Give two examples.

17.3 Fourier Series of a Periodic Function

17.3.1. Fourier series. Recall that if $f(x)$ is infinitely differentiable at a point $x = a$, then it has a Taylor series about that point,

$$\text{TS}\, f\Big|_{x=a} = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{1}$$

For that series to be useful, we need two things. First, we need it to converge on some interval I. Second, if it does converge on I, then we need its sum function to be the same as the original function $f(x)$.* If that is the case, then we say that the Taylor series $\text{TS}\, f\Big|_{x=a}$ **represents** f on I, and we can write

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{2}$$

on I. We know from experience that Taylor series representations are useful in many ways.

Similarly, there are other types of representations that are useful as well. In this chapter we are concerned with the representation of periodic functions, not by Taylor series but by **trigonometric series**, that is, by series of the form

$$a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right). \tag{3}$$

First, observe that *(3) is a periodic function with fundamental period 2ℓ*. To see this, note that (3) is a linear combination of the functions 1, $\cos(\pi x/\ell)$, $\sin(\pi x/\ell)$,

*For instance, we saw in Example 2 of Section 13.5 that it is possible for $\text{TS}\, f\Big|_{x=a}$ to converge, but not to $f(x)$.

$\cos(2\pi x/\ell)$, $\sin(2\pi x/\ell)$, and so on, and that these functions are periodic with the following periods:

$$
\begin{aligned}
&& 1:&\quad \text{arbitrary}, && \text{(4a)}\\
\cos\frac{\pi x}{\ell} &\quad\text{and}\quad \sin\frac{\pi x}{\ell}: &&\quad \underline{2\ell}, 4\ell, 6\ell, 8\ell, \dots, && \text{(4b)}\\
\cos\frac{2\pi x}{\ell} &\quad\text{and}\quad \sin\frac{2\pi x}{\ell}: &&\quad \ell, \underline{2\ell}, 3\ell, 4\ell, \dots, && \text{(4c)}\\
\cos\frac{3\pi x}{\ell} &\quad\text{and}\quad \sin\frac{3\pi x}{\ell}: &&\quad \frac{2\ell}{3}, \frac{4\ell}{3}, \underline{2\ell}, \frac{8\ell}{3}, \dots, && \text{(4d)}
\end{aligned}
$$

and so on. The smallest period shared by all the terms is 2ℓ [underlined in (4)], so the fundamental period of (3) is 2ℓ.* Thus, perhaps the trigonometric series (3) can be used to represent periodic functions of period 2ℓ. (When we say of period 2ℓ, we shall mean of fundamental period 2ℓ.)

Specifically, if $f(x)$ is periodic, of period 2ℓ, then we define the **Fourier series of f**, say FS f, as

$$\boxed{\text{FS } f = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right),} \tag{5a}$$

where the coefficients are given by the **Euler formulas**

$$\boxed{\begin{aligned}
a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx, &&\\
a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, && n = 1, 2, \dots\\
b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx, && n = 1, 2, \dots
\end{aligned}} \tag{5b,c,d}$$

and are known as the **Fourier coefficients** of f.

For FS f to represent f we need the series to converge, and we need its sum function to be the same as the original function $f(x)$. Various theorems are available, that give sufficient conditions on f for FS f to represent f. One such theorem, that is easily applied and which covers the vast majority of periodic functions that arise in applications, is as follows.†

THEOREM 17.3.1 *Fourier Convergence Theorem*
Let f be 2ℓ-periodic, and let f and f' be piecewise continuous on $[-\ell, \ell]$. Then

*Actually, the fundamental period of the function (3) *may* be less than 2ℓ. For example, if all the coefficients except for a_3 are zero, then the fundamental period is $2\ell/3$. Thus, we should say that (3) is always periodic with period 2ℓ; its fundamental period is at most 2ℓ, but may be less.

†For proof of a slightly stronger version of Theorem 17.3.1, see R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed. (New York: McGraw-Hill, 1978), Sec.41.

the Fourier series given by (5) converges to $f(x)$ at every point x at which f is continuous, and to the mean value $[f(x+) + f(x-)]/2$ at every point x at which f is discontinuous.

Piecewise continuity is defined in Section 5.2. By $f(x+)$ and $f(x-)$ we mean the **right-** and **left-hand limits** of f,

$$f(x+) \equiv \lim_{h \to 0} f(x+h) \qquad \text{and} \qquad f(x-) \equiv \lim_{h \to 0} f(x-h), \tag{6}$$

where $h \to 0$ through positive values. [If $f(x+) = f(x-) = f(x)$, then f is continuous at x, otherwise it is discontinuous there.]

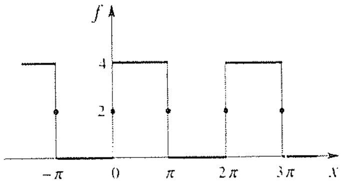


Figure 1. Square wave.

EXAMPLE 1. *Square Wave.* Consider the "square wave" f shown in Fig. 1, where $f(x)$ is defined as 2 at $x = 0, \pm\pi, \pm 2\pi, \ldots$, as indicated by the heavy dots. Since the period referred to in Theorem 17.3.1 and in (5) is 2ℓ, and the period is seen from Fig. 1 to be 2π, it follows that $\ell = \pi$. Both f and f' are piecewise continuous on $[-\pi, \pi]$, so the theorem applies. Let us use (5) to work out the Fourier series of f and examine its convergence to the square wave f using computer plots of the partial sums of the series.

First, (5b) gives

$$\begin{aligned} a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx \\ &= \frac{1}{2\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\,dx\right] = 2. \end{aligned} \tag{7}$$

Actually, $a_0 = 2$ could have been seen by inspection because the right-hand side of (5b) is, by definition, the **average value** of f over one period and, from Fig. 1, we can see that for our square wave that average value is 2.

Next,

$$\begin{aligned} a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \\ &= \frac{1}{\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\cos nx\,dx\right] = \left.\frac{4\sin nx}{n\pi}\right|_{0}^{\pi} = 0 \end{aligned} \tag{8}$$

since $\sin n\pi = 0$ and $\sin 0 = 0$. Finally,

$$\begin{aligned} b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \\ &= \frac{1}{\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\sin nx\,dx\right] = \frac{4}{n\pi}(1-\cos n\pi). \end{aligned} \tag{9}$$

We can leave (9) as it stands, but it is best to note that $\cos n\pi = (-1)^n$, so that we can reduce (9) to

$$b_n = \frac{4}{n\pi}[1-(-1)^n] = \begin{cases} \dfrac{8}{n\pi}, & n = 1, 3, \ldots \\ 0, & n = 2, 4, \ldots. \end{cases} \tag{10}$$

Thus the Fourier series representation of f is

$$f(x) = 2 + \frac{8}{\pi} \sum_{1,3,\dots}^{\infty} \frac{1}{n} \sin nx, \tag{11}$$

where "1, 3, . . ." tells us to omit the terms corresponding to $n = 2, 4, \dots$. If preferred, (11) can be expressed, equivalently, as

$$f(x) = 2 + \frac{8}{\pi} \sum_{1}^{\infty} \frac{\sin (2n-1)x}{2n-1}, \tag{12}$$

with $\sum_1^\infty$ denoting the usual summation over $n = 1, 2, 3, \dots$.

Observe from Fig. 1 that the defined values of $f(x)$ at the jump discontinuities ($x = n\pi$) coincide with the mean values $[f(x+) + f(x-)]/2 = 2$, so Theorem 17.3.1 assures us that (12) is indeed an equality for every value of x.

Is it not remarkable that a linear combination of sines and cosines, each of which is beautifully smooth (infinitely differentiable for all x, i.e., of class C^∞), can sum to a function with jump discontinuities? To obtain insight as to how this convergence to f is accomplished, it will be illuminating to plot some of the partial sums of the series. Writing out (12) as

$$f(x) = 2 + \frac{8}{\pi} \sin x + \frac{8}{3\pi} \sin 3x + \frac{8}{5\pi} \sin 5x + \cdots, \tag{13}$$

we define the partial sums of the series as

$$\begin{aligned} s_1(x) &= 2, \\ s_2(x) &= 2 + \frac{8}{\pi} \sin x, \\ s_3(x) &= 2 + \frac{8}{\pi} \sin x + \frac{8}{3\pi} \sin 3x, \end{aligned}$$

and so on: that is, $s_N(x)$ is the sum of the first N terms.

We see from Fig. 2 that although the first partial sum $s_1(x)$ is merely a constant it "does the best it can" by equalling the average value of f. If we are to provide the needed correction to $s_1(x)$ we need to pull it up on the right and push it down on the left, as

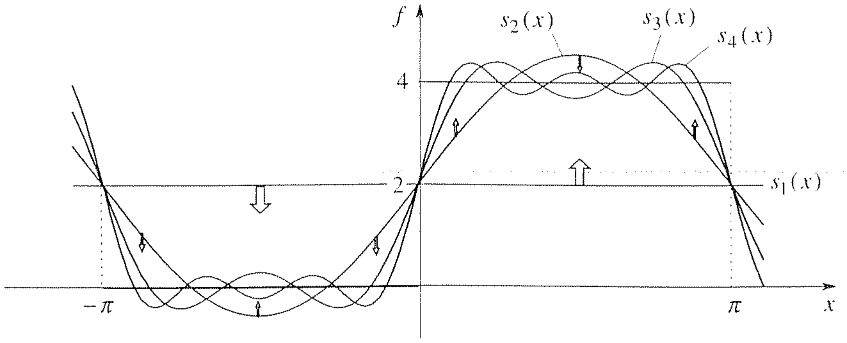


Figure 2. The first few partial sums.

indicated by the broad arrows in the figure. A $\sin x$ term can accomplish that (a $\cos x$ term is not appropriate since we need an antisymmetric correction, not a symmetric one) and an amplitude a bit greater than 2 seems optimal. That correction is precisely what is provided by the second term in the series, so $s_2(x)$ begins to take on the desired shape. To correct $s_2(x)$, in turn, we need to push it up and down according to the six thin arrows. Such correction dictates the need for a third sine harmonic, and such a term is indeed forthcoming as the third term in the series. By adding more and more terms of the series, the graph comes closer and closer to the square wave f. Observe also that the graph of every partial sum passes through the mean value 2 at each jump discontinuity, which result is also clear from (12) since $\sin n\pi = 0$ for each n.

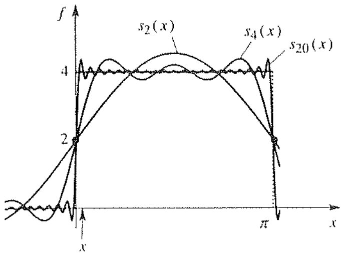


Figure 3. Gibbs phenomenon.

Next, consider the graph of $s_N(x)$ for larger N, for instance for $N = 20$ (Fig. 3). As x increases from zero, $s_{20}(x)$ rises sharply from the mean value 2, overshoots the value 4, and settles down close to 4 – until x approaches π, where the same sort of overshoot is followed by a steep descent to the mean value 2 at $x = \pi$. By periodicity, these results repeat over each period.

Strangely, the overshoot does not diminish as N increases. For instance, the peak values of $s_2(x)$, $s_4(x)$, and $s_{20}(x)$ are 4.548, 4.376, and 4.360, respectively, and it can be shown that the overshoot approaches a limiting value of around 9% of the jump (the jump is 4, so 9% of 4 is 0.36) as $N \to \infty$. This persistent 9% overshoot occurs not only in this example but in the Fourier series representation of any function with a jump discontinuity, and is known as the **Gibbs phenomenon.**[*]

In view of this overshoot, one may well wonder how convergence to the square wave is attained, for are there not always x locations at which the error is around 9% no matter how many terms are summed? The key point is that when we say that $\lim_{N\to\infty} s_N(x) = f(x)$ it is N that is varying; x is *fixed.* Picking any x point (Fig. 3), as close to the origin as we like, as N is increased the overshoot "spike" eventually moves to the left of x, and subsequent values of $s_N(x)$ do settle down and converge to $f(x)$. Nonetheless, the Gibbs phenomenon is of great practical importance because it implies that the convergence of Fourier series may be painfully slow (and expensive) in the vicinity of a jump discontinuity.[†]

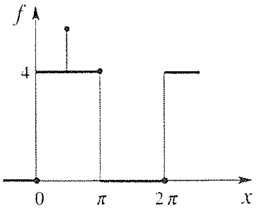


Figure 4. Modified f.

COMMENT 1. We have already noted that the Fourier series of the square wave $f(x)$ shown in Fig. 1 converges to $f(x)$ at every point x. Suppose, instead, we define f as shown in Fig. 4. That is, the modified f is 0 at $x = 0, \pm 2\pi, \pm 4\pi, \dots$, it is 4 at $x = \pm\pi, \pm 3\pi, \dots$, and it is 10^6 (not shown to scale) at $x = \frac{\pi}{2}, -\frac{3\pi}{2}, \frac{5\pi}{2}, -\frac{7\pi}{2}, \dots$. The Fourier series of the modified f, say f_{mod}, will be identical to the Fourier series of f, because the Fourier coefficients are computed, according to (5b,c,d), by integrals, and changing the value of the integrands at a number of isolated points does not change the value of those integrals.

[*] *Josiah Willard Gibbs* (1839–1903) was one of the first important American mathematical physicists, and is especially well known for his work on vector analysis and thermodynamics. The Gibbs phenomenon was noted by Gibbs in 1899 (*Nature*, Vol. 59, p. 606) and explained in 1906 by Maxime Bôcher (*Annals of Mathematics*, Vol. 2, No. 7, p. 81).

[†] When studying infinite series in the calculus one learns to classify series as convergent or divergent, with the implication that convergent series are "good" and divergent series are "bad." Practically speaking, however, a series may converge – but so slowly as to be almost worthless. Thus, interest exists in developing *acceleration* techniques that permit us to transform a given slowly convergent series into a more rapidly convergent one. For a numerical technique to suppress the Gibbs phenomenon, see Forman Acton, *Numerical Methods That Work* (Washington, D.C.: Mathematical Assn. of America, 1990).

Thus, the Fourier series of f_{mod} will converge to $f_{\mathrm{mod}}(x)$ for every point x except points such as $x = 0, \frac{\pi}{2}$, and π, where it will converge to 2, 4, and 2, respectively. Such pointwise discrepancies will cause no problem in applications. Thus, from this point forward (except in the exercises to this section) we will not bother to show the heavy dots, as we did in Figs. 1 and 4, and will simply show these graphs as in Fig. 5. ▮

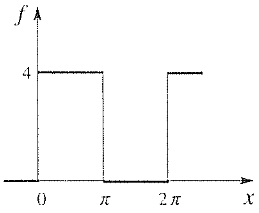


Figure 5. No circles or dots.

Thus far, the concepts of even and odd functions, introduced in Section 17.2, have not been used here. Suppose now that f is an *even* 2ℓ-periodic function. Then the integrands $f(x)$ and $f(x)\cos(n\pi x/\ell)$ in (5b) and (5c) are even functions, and the integrand $f(x)\sin(n\pi x/\ell)$ in (5c) is odd, so these Euler formulas simplify to

$$\begin{aligned} a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = \frac{1}{\ell}\int_0^{\ell} f(x)\,dx, \\ a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{2}{\ell}\int_0^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, \\ b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = 0. \end{aligned} \tag{14}$$

Observe that only the constant and cosine terms survive, there being no need for sine terms in representing an even function. Similarly, if f is *odd*, then

$$\begin{aligned} a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = 0, \\ a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = 0, \\ b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = \frac{2}{\ell}\int_0^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx, \end{aligned} \tag{15}$$

and the constant and cosine terms drop out.

More generally, f is neither even nor odd. The constant and cosine terms in the Fourier series of $f(x)$ represent the even part $f_e(x)$, and the sine terms represent the odd part $f_o(x)$. That is,

$$\begin{aligned} f(x) &= \left[a_0 + \sum_1^{\infty} a_n \cos\frac{n\pi x}{\ell}\right] + \left[\sum_1^{\infty} b_n \sin\frac{n\pi x}{\ell}\right] \\ &= \qquad\quad f_e(x) \qquad\quad + \qquad f_o(x), \end{aligned} \tag{16}$$

and

$$f_e(x) = a_0 + \sum_1^{\infty} a_n \cos\frac{n\pi x}{\ell}, \tag{17a}$$

$$f_o(x) = \sum_1^{\infty} b_n \sin\frac{n\pi x}{\ell}. \tag{17b}$$

EXAMPLE 2. Let us work out the Fourier series of the periodic function f, the graph of which is given in Fig. 6. Its period is 16, so $\ell = 8$. We can compute a_0 by (5b), but we can see from the figure (dividing the net area over one period by 16) that the average height is $-\frac{1}{4}$ so $a_0 = -\frac{1}{4}$. Further, we see that f is even, in which case (14) gives $b_n = 0$ and

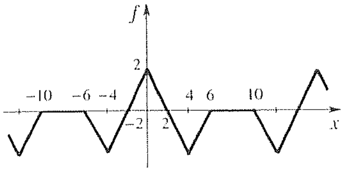


Figure 6. f in Example 2.

$$
\begin{aligned}
a_n &= \frac{2}{\ell}\int_0^\ell f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{2}{8}\int_0^8 f(x)\cos\frac{n\pi x}{8}\,dx \\
&= \frac{1}{4}\left[\int_0^4 (2-x)\cos\frac{n\pi x}{8}\,dx + \int_4^6 (x-6)\cos\frac{n\pi x}{8}\,dx + \int_6^8 0\cos\frac{n\pi x}{8}\,dx\right] \\
&= \frac{1}{4}\left[2\frac{\sin(n\pi x/8)}{n\pi/8} - \frac{1}{(n\pi/8)^2}\left(\cos\frac{n\pi x}{8} + \frac{n\pi x}{8}\sin\frac{n\pi x}{8}\right)\right]\Bigg|_0^4 \\
&\quad + \frac{1}{4}\left[\frac{1}{(n\pi/8)^2}\left(\cos\frac{n\pi x}{8} + \frac{n\pi x}{8}\sin\frac{n\pi x}{8}\right) - 6\frac{\sin(n\pi x/8)}{n\pi/8}\right]\Bigg|_4^6 \\
&= \frac{16}{n^2\pi^2}\left(1 - 2\cos\frac{n\pi}{2} + \cos\frac{3n\pi}{4}\right).
\end{aligned} \tag{18}
$$

Thus,

$$
f(x) = -\frac{1}{4} + \frac{16}{\pi^2}\sum_{n=1}^{\infty}\frac{1 - 2\cos\frac{n\pi}{2} + \cos\frac{3n\pi}{4}}{n^2}\cos\frac{n\pi x}{8}, \tag{19}
$$

with the equality holding for all x, without exception.

COMMENT. In Example 1 we simplified $\sin n\pi$ and $\cos n\pi$ as 0 and $(-1)^n$, respectively, but the $\cos(n\pi/2)$ and $\cos(3n\pi/4)$ terms in (19) are trickier (depending on n, $\cos(n\pi/2)$ takes on the values $0, \pm 1$, and $\cos(3n\pi/4)$ takes on the values $0, \pm 1, \pm\sqrt{2}/2$), so we will leave them intact. ▮

Recall from the calculus that the ***p*-series**,

$$
\sum_1^\infty \frac{1}{n^p}, \tag{20}
$$

converges if $p > 1$ and diverges if $p \leq 1$, the borderline divergent case of $p = 1$ corresponding to the **harmonic series**

$$
\sum_1^\infty \frac{1}{n}. \tag{21}
$$

Further, you may recall that the *alternating* harmonic series

$$
\sum_1^\infty \frac{(-1)^n}{n} \tag{22}
$$

is convergent – barely convergent in that it converges extremely slowly,[*] and is only conditionally convergent rather than absolutely convergent. The Fourier series (11) is similar to the alternating series (22) and, like (22), converges "by the skin of its teeth,"[†] which result should not be surprising since (11) amounts to the representation of a function having jump discontinuities in terms of a collection of infinitely smooth sines. The function f in Example 2 is better behaved in that it is continuous and, sure enough, the coefficients in (19) die out faster, proportional to $1/n^2$. That pattern persists: as we consider periodic functions that are better and better behaved ($C^0, C^1, C^2, \dots$) we find that their Fourier coefficients die out (as $n \to \infty$) more and more rapidly. In general, if a periodic function is C^N, we can expect its Fourier coefficients to tend to zero at least as fast as $1/n^{N+2}$.[‡] As an extreme case, observe that the periodic function $f(x) = 5\sin 3x$, say, is infinitely smooth (i.e., it is of class C^∞), and its Fourier coefficients do tend to zero infinitely fast. That is, its Fourier series is simply one term, FS $\{5\sin 3x\} = 5\sin 3x$.

17.3.2. Euler's formulas. We chose to state the Fourier convergence theorem and to move quickly into examples to solidify the Fourier series concept. Now let us back up and show where Euler's formulas (5b,c,d) come from. Accepting that a (sufficiently well-behaved) 2ℓ-periodic function f can be represented in the Fourier series form

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right), \tag{23}$$

we focus attention on how to compute the coefficients a_0, a_n, and b_n. We will need the following elementary integral formulas:

$$\int_{-\ell}^{\ell} \cos \frac{m\pi x}{\ell} \cos \frac{n\pi x}{\ell}\, dx = \begin{cases} 0 & m \neq n \\ \ell & m = n \neq 0 \\ 2\ell & m = n = 0 \end{cases} \tag{24a}$$

$$\int_{-\ell}^{\ell} \sin \frac{m\pi x}{\ell} \sin \frac{n\pi x}{\ell}\, dx = \begin{cases} 0 & m \neq n \\ \ell & m = n \neq 0 \end{cases} \tag{24b}$$

$$\int_{-\ell}^{\ell} \cos \frac{m\pi x}{\ell} \sin \frac{n\pi x}{\ell}\, dx = 0 \qquad \text{for all } m, n, \tag{24c}$$

where m and n are integers. The three zero results in (24) are due to cancellation of positive and negative areas; the two nonzero results, in (24a) and (24b), occur when $m = n$, in which case the integrands are squared quantities and no such cancellation can occur.

[*] For three-significant-figure accuracy we need to sum around 10^3 terms; for six-significant-figure accuracy we need around 10^6 terms, and so on.

[†] The $1/n$ decay, in (22) and (11), is not enough to induce convergence. Rather, these series converge (according to the *Dirichlet test*) because $(-1)^n$ and $\sin nx$ have *bounded partial sums*; that is, there exist finite numbers A and B such that $|\sum_{n=1}^{N}(-1)^n| < A$ and $|\sum_{n=1}^{N} \sin nx| < B$ for all N's, no matter how large.

[‡] See, for instance, H. S. Carslaw, *Fourier Series* (New York: Dover, 1930), pp. 269–271.

For definiteness, let us solve (23) for a_2. To do so, multiply both sides of (23) by $\cos(2\pi x/\ell)$ and integrate over one period. That step gives

$$\begin{aligned}\int_{-\ell}^{\ell} f(x)\cos\frac{2\pi x}{\ell}\,dx &= \int_{-\ell}^{\ell}\left[a_0+\sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]\cos\frac{2\pi x}{\ell}\,dx\\ &= a_0\int_{-\ell}^{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_1\int_{-\ell}^{\ell}\cos\frac{\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\ &\quad +b_1\int_{-\ell}^{\ell}\sin\frac{\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_2\int_{-\ell}^{\ell}\cos\frac{2\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\ &\quad +b_2\int_{-\ell}^{\ell}\sin\frac{2\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_3\int_{-\ell}^{\ell}\cos\frac{3\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\ &\quad +\cdots\\ &= 0+0+0+a_2\ell+0+0+\cdots\\ &= a_2\ell, \end{aligned}\tag{25}$$

so

$$a_2 = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{2\pi x}{\ell}\,dx, \tag{26}$$

where the zeros following the third equal sign follow from (24a,b,c), and the $a_2\ell$ term follows from (24a) with $m=n=2$. [The first zero follows from (24a) with $m=0$ and $n=2$.] More generally, to solve for a_n multiply (23) by $\cos(n\pi x/\ell)$ and integrate over one period; to solve for b_n multiply by $\sin(n\pi x/\ell)$ instead, and to solve for a_0 multiply by 1 instead. These steps, with the help of (24), give the Euler formulas (5b,c,d).

The interchange in the order of summation and integration, in the second step in (25) needs justification, but we will postpone discussion of such technical points until Section 17.5.

There is a strong analogy between the calculation of the coefficients c_n in the expansion

$$\mathbf{v} = \sum_{n=1}^{N} c_n\mathbf{e}_n \tag{27}$$

of a vector $\mathbf{v}$ in an N-dimensional vector space, in terms of an orthogonal basis $\{\mathbf{e}_1,\dots,\mathbf{e}_N\}$. To solve for c_2, for instance, we take advantage of the orthogonality of the basis vectors, and dot both sides with $\mathbf{e}_2$. That step gives

$$\begin{aligned}\mathbf{v}\cdot\mathbf{e}_2 &= c_1\mathbf{e}_1\cdot\mathbf{e}_2 + c_2\mathbf{e}_2\cdot\mathbf{e}_2 + c_3\mathbf{e}_3\cdot\mathbf{e}_2+\cdots+c_N\mathbf{e}_N\cdot\mathbf{e}_2\\ &= 0 + c_2\mathbf{e}_2\cdot\mathbf{e}_2 + 0+\cdots+0\\ &= c_2\mathbf{e}_2\cdot\mathbf{e}_2,\end{aligned}\tag{28}$$

so

$$c_2 = \frac{\mathbf{v}\cdot\mathbf{e}_2}{\mathbf{e}_2\cdot\mathbf{e}_2} \tag{29}$$

and similarly for $c_1, c_3, c_4, \dots, c_N$. In fact, in Sections 17.6 and 17.7 we will take another look at Fourier series from exactly that point of view. There, we will understand (23) as the expansion of a vector f in an infinite-dimensional function space in terms of the base vectors $1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \dots,$ which are orthogonal by virtue of equations (24); in fact, (24) are often called **orthogonality relations**.

17.3.3. Applications. Fourier series are indispensable in our study of PDE's (partial differential equations) in the next three chapters. Here, we give two applications to physical systems that are governed by ODE's and subjected to periodic forcing functions. Our approach will be *formal*, by which we mean that although we believe the results to be correct, we will not rigorously justify all of the steps involved in their derivation.

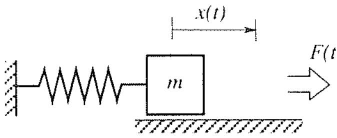


Figure 7. Forced mechanical oscillator.

EXAMPLE 3. *Periodically Driven Oscillator.* Consider the driven mechanical oscillator shown in Fig. 7 and governed by the differential equation of motion

$$mx'' + cx' + kx = F(t), \tag{30}$$

where m, c, k are the mass, damping coefficient (associated with some combination of viscous damping due to a film of lubricating oil between the mass and the table, and air resistance), and spring stiffness, respectively. Let $m = 1$ kg, $c = 0.04$ kg/sec, and $k = 15$ kg/sec^2, and let $F(t)$ (in newtons) be as shown in Fig. 8. F consists of an endless sequence of pulses, each having unit area (except the first, which is only a "half pulse"). In mechanics, $\int_{t_1}^{t_2} F(t)\,dt$ is the **impulse** delivered by the force F between times t_1 and t_2, so F consists of a periodic sequence of unit impulses, of period 2π. Thus, $\ell = \pi$ in (5). Even though the starting time is $t = 0$, so $t \geq 0$, we can think of F as the extended function shown in Fig. 9, which is even. Thus, if we expand F in a Fourier series we have, for its coefficients, $a_0 =$ average value $= 1/(2\pi)$, $b_n = 0$ because F is even, and, from (14),

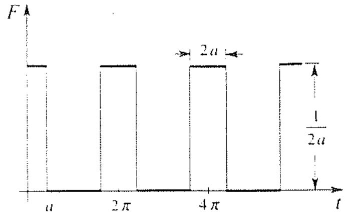


Figure 8. Forcing function F.

$$a_n = \frac{2}{\pi}\int_0^{\pi} F(t)\,\cos nt\,dt = \frac{2}{\pi}\int_0^{a} \frac{1}{2a}\cos nt\,dt = \frac{\sin na}{n\pi a}. \tag{31}$$

Thus, (30) becomes

$$x'' + 0.04x' + 15x = \frac{1}{2\pi} + \sum_1^{\infty} \frac{\sin na}{n\pi a}\cos nt. \tag{32}$$

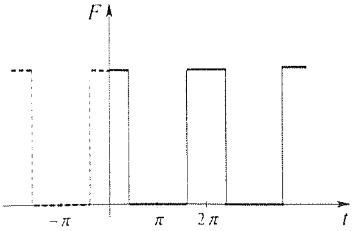


Figure 9. F is even.

Recall, from the general theory in Chapter 3, that if

$$L[x] = f_1 + \cdots + f_N, \tag{33}$$

where L is shorthand for a linear second-order differential operator, then the general solution of (33) is of the form

$$x(t) = C_1 x_{h1}(t) + C_2 x_{h2}(t) + x_{p1}(t) + \cdots + x_{pN}(t), \tag{34}$$

where C_1, C_2 are arbitrary constants, x_{h1} and x_{h2} are linearly independent homogeneous solutions, and $x_{p1}, \dots, x_{pN}$ are particular solutions corresponding to the forcing functions

$f_1, \dots, f_N$, respectively. If two initial conditions are prescribed, they enable us to determine C_1 and C_2.

In the present example x_{h1} and x_{h2} are oscillatory, but with a slow exponential decay due to the $0.04x'$ damping term in (32).* As $t \to \infty$ those terms decay to zero, leaving us with the particular solution as the steady-state response. Let us limit our attention to finding, and discussing, the steady-state response.

Our plan is to find a particular solution corresponding to each forcing term on the right side of (32), using the method of undetermined coefficients, and then to superimpose those solutions, as in (34). Consider

$$x'' + 0.04x' + 15x = \cos nt, \tag{35}$$

and seek x_p in the form $A \cos nt + B \sin nt$. Putting that form into the left-hand side of (35) enables us to solve for A and B, and we obtain

$$x_p(t) = \frac{15 - n^2}{(15 - n^2)^2 + 0.0016n^2} \cos nt + \frac{0.04n}{(15 - n^2)^2 + 0.0016n^2} \sin nt. \tag{36}$$

Further, we find that a particular solution corresponding to the $1/2\pi$ forcing term in (32) is $x_p(t) = 1/30\pi$, so the desired steady-state response is, by linearity and superposition,

$$\begin{aligned} x(t) = \frac{1}{30\pi} + \sum_1^\infty \frac{\sin na}{n\pi a} \bigg[& \frac{15 - n^2}{(15 - n^2)^2 + 0.0016n^2} \cos nt \\ & + \frac{0.04n}{(15 - n^2)^2 + 0.0016n^2} \sin nt \bigg]. \end{aligned} \tag{37}$$

Choosing a specific value for a, we can use (37) to compute $x(t)$. However, for purposes of understanding it will be useful to express the square-bracketed term in (37), equivalently, in the form $A_n \cos(nt + \phi_n)$, where the amplitude A_n and the phase ϕ_n are

$$A_n = \frac{1}{\sqrt{(15 - n^2)^2 + 0.0016n^2}} \qquad \text{and} \qquad \phi_n = \tan^{-1}\left(\frac{0.04n}{n^2 - 15}\right). \tag{38}$$

Let $a = \pi/8$, say. Then, using (38) we can write out (37) as

$$\begin{aligned} x(t) &= 0.0106 + 0.0222 \cos(t - 0.0029) + 0.0261 \cos(2t - 0.0073) \\ &\quad + 0.0416 \cos(3t - 0.0200) + 0.2001 \cos(4t + 3.3002) \\ &\quad + 0.0150 \cos(5t + 3.1616) + \cdots. \end{aligned} \tag{39}$$

Consider the terms on the right-hand side of (39). The first, 0.0106, is the response to the $1/2\pi$ forcing term in (32), the second is the response to the $n = 1$ term in (32), the third is the response to the $n = 2$ term in (32), and so on. Observe that if we ignore the $0.04x'$ damping term in (35) then the natural frequency of the oscillator is $\sqrt{15}$ and that the $\cos nt$ forcing function comes close to that natural frequency when $n = 4$. Sure enough, the $n = 4$ term in (39) has by far the largest amplitude, 0.2001. Observe further that the phase ϕ_n is rather small for $n < 4$. That is, the response term is almost in phase with the forcing term

*If this sounds unfamiliar we urge you to review Section 3.5.

– because the damping is light. However, for $n \geq 4$ the ϕ_n's are approximately π. That is, when the frequency of the forcing terms exceeds the approximate natural frequency $\sqrt{15}$ the phase increases to around π, so that the response is around 180° out of phase with the forcing function.*

Plotting the steady-state response $x(t)$, given by (39), in Fig. 10, we can see how the response is dominated by the $0.2001 \cos(4t + 3.3002)$ term, as discussed above. Observe that it suffices to plot $x(t)$ over any 2π interval since it is 2π-periodic.

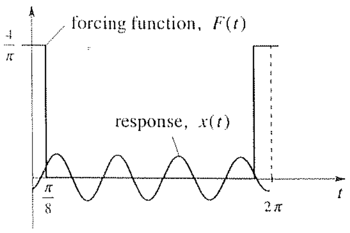


Figure 10. Steady-state response.

COMMENT 1. Why did we expand $F(t)$ in a Fourier series? That step gave us $F(t)$ as a linear combination of elementary functions, the response to each of which was readily found, with the total response then built up by superposition. Alternatively, the Laplace transform method would also have been convenient. (See Exercise 17 for other ideas.)

COMMENT 2. We stated that our solution would be formal rather than rigorous. The point that we did not justify is as follows. The expression (34) satisfies (33) because

$$\begin{aligned} L[x] &= L\left[C_1 x_{h1} + C_1 x_{h2} + \sum_1^N x_{pn}\right] \\ &= C_1 L[x_{h1}] + C_2 L[x_{h2}] + L\left[\sum_1^N x_{pn}\right] \\ &= 0 + 0 + \sum_1^N L\left[x_{pn}\right] = \sum_1^N f_n. \end{aligned} \tag{40}$$

However, in the present case $N = \infty$, so the step

$$L\left[\sum_1^\infty x_{pn}\right] = \sum_1^\infty L\left[x_{pn}\right] \tag{41}$$

amounts to an interchange in the order of two limit operations, the derivatives in L and the infinite series. The validity of such interchange will not be covered until Section 17.5.

COMMENT 3. Two limiting cases are of interest (and are available to us because we left a as a parameter). As $a \to \pi$, $F(t)$ tends to the constant $1/2\pi$. In that case the series in (32) vanishes and the steady-state response is simply $x(t) = 1/30\pi$. More subtle is the case where $a \to 0$ since then $F(t)$ becomes a sequence of hammer blows, each imparting unit impulse. Discussion of this case is left for the exercises.

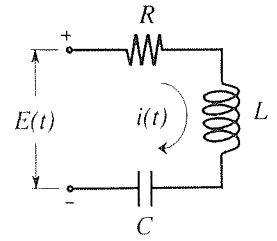


Figure 11. Electrical analog.

COMMENT 4. The electrical analog of the forced mechanical oscillator, governed by (30) and shown in Fig. 7, is the equation

$$LQ'' + RQ' + \frac{1}{C}Q = E(t) \tag{42}$$

governing the circuit shown in Fig. 11, where $Q(t)$ is the charge on the capacitor. ▮

*See Fig. 2 in Section 3.8. To understand the present example you may need to review both Sections 3.5 and 3.8.

EXAMPLE 4. *Infinite Beam on Elastic Foundation.* Consider an infinitely long beam on an elastic foundation, sketched in Fig. 12. The constant k is called the *foundation modulus* (i.e., the spring stiffness per unit x-length) and the square wave $w(x)$ is a prescribed periodic loading (force per unit length).[*] Physically, the beam might be a train track, with the elastic foundation used to model the track bed. We wish to determine the vertical deflection $u(x)$.

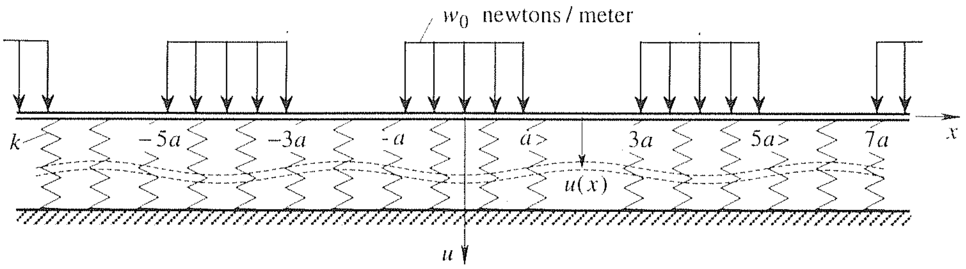


Figure 12. Infinite beam on elastic foundation.

According to the classical *Euler beam theory*[†] the deflection $u(x)$ resulting from a load distribution $p(x)$ newtons per meter satisfies the fourth-order differential equation

$$EIu'''' = p(x), \tag{43}$$

where E and I are physical constants of the beam; EI is called the *flexural rigidity* of the beam (and is considered here as a known constant) since u'''', and hence the deflection u, is inversely proportional to EI. Now $p(x)$ is the *net* loading and consists of the applied periodic loading $w(x)$ downward and the spring force $ku(x)$ upward. (We neglect the weight of the beam, for simplicity.) Thus, $p(x) = w(x) - ku(x)$, and (43) becomes

$$EIu'''' + ku = w(x). \tag{44}$$

This problem is similar to the preceding one in that they both involve differential equations with periodic forcing functions, but in this case the independent variable is x and the interval is $-\infty < x < \infty$. As in Example 3, we begin by expanding the forcing function in a Fourier series,

$$w(x) = \frac{w_0}{2} + \frac{2w_0}{\pi} \sum_{n=1}^{\infty} \frac{\sin(n\pi/2)}{n} \cos\frac{n\pi x}{2a}. \tag{45}$$

Rather than find the response to each forcing term and adding them up, let us use a slightly different procedure (which could have been used in Example 3 as well). Namely, anticipating that $u(x)$ will be an even periodic function, of the same period as $w(x)$ (i.e., 4a), let us seek $u(x)$ in the form

$$u(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{2a}. \tag{46}$$

[*] That is, the downward load between any points x_1 and x_2 ($> x_1$) is $\int_{x_1}^{x_2} w(x)\,dx$.

[†] See, for example, S. Timoshenko, *Strength of Materials*, Part I (Princeton, NJ: D. Van Nostrand, 1955).

Formally differentiating (46) termwise four times, putting this result and (45) into (44), and formally equating coefficients of each cosine harmonic gives

$$ka_0 = \frac{w_0}{2} \tag{47}$$

and

$$\left(EI\,\frac{n^4\pi^4}{16a^4} + k\right) a_n = \frac{2w_0}{\pi}\,\frac{\sin(n\pi/2)}{n}, \tag{48}$$

so

$$a_0 = \frac{w_0}{2k}, \qquad a_n = \frac{2w_0}{\pi}\,\frac{\sin(n\pi/2)}{n[EI\,(n^4\pi^4/16a^4) + k]}. \quad (n \geq 1) \tag{49}$$

Thus we have the formal solution

$$u(x) = \frac{w_0}{2k} + \frac{32w_0a^4}{\pi}\sum_{n=1}^{\infty}\frac{\sin(n\pi/2)}{n(EIn^4\pi^4 + 16a^4k)}\cos\frac{n\pi x}{2a}. \tag{50}$$

COMMENT 1. It is striking that the terms in the series die out so rapidly, proportional to $1/n^5$ as $n \to \infty$, so that we can expect merely the first couple of terms of (50) to give a good approximation to $u(x)$,

$$u(x) \approx \frac{w_0}{2k} + \frac{32w_0a^4}{\pi(\pi^4EI + 16a^4k)}\cos\frac{\pi x}{2a}. \tag{51}$$

How are we to understand the input $w(x)$ being discontinuous and the output $u(x)$ being quite smooth? Physically, we don't expect $u(x)$ to be discontinuous or "kinky" just because the loading $w(x)$ is, because a train track is too "stiff" for that. Mathematically, observe that in essence (though not procedurally, unless $k = 0$) we solve (44) for u by four integrations of w. Now, what happens when we repeatedly integrate a discontinuous function? Consider a Heaviside step function, for simplicity, in place of w. Integrating from $-\infty$, say, to a variable point x, gives

$$\begin{aligned}\int_{-\infty}^{x} H(\xi)\,d\xi &= \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases} \\ &= xH(x),\end{aligned} \tag{52}$$

which is a "ramp" function. Integrating $xH(x)$, in turn, gives $(x^2/2)H(x)$; integrating $(x^2/2)H(x)$ gives $(x^3/6)H(x)$, and so on, as displayed in Fig. 13. That is, *integration is a smoothing operation*: $H(x)$ is discontinuous, its integral $xH(x)$ is C^0 (continuous), the integral of the latter is C^1, the integral of the latter is C^2, and so on.[*] From this rough argument, we expect the response $u(x)$ to the discontinuous load $w(x)$ to be C^3, so its Fourier coefficients should tend to zero like $1/n^5$, and that is precisely what is revealed by (50).

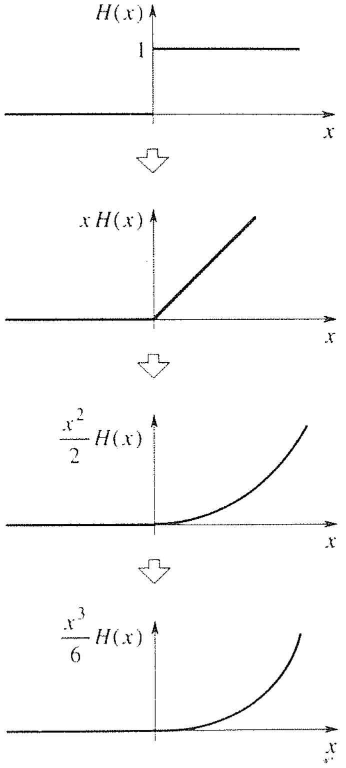


Figure 13. The smoothing effect of integration.

COMMENT 2. Actually, (50) is a *particular* solution of (44). To obtain the general solution we need to add the homogeneous solution

$$e^{\beta x}(A\sin\beta x + B\sin\beta x) + e^{-\beta x}(C\sin\beta x + D\sin\beta x), \tag{53}$$

[*] Conversely, differentiation has the opposite effect. For instance, $(x^3/6)H(x)$ is smooth but has a singular behavior at the origin, that is brought to light by repeated differentiations.

where $\beta \equiv \sqrt[4]{k/EI}/\sqrt{2}$. To determine A, B, C, D we need some sort of boundary conditions at the ends of the interval, that is, as $x \to \pm\infty$. By way of such conditions, let us require u to be *bounded* as $x \to \pm\infty$, as would be a reasonable requirement for a loaded train track. Then the behavior $e^{\beta x} \to \infty$ as $x \to +\infty$ implies that we need $A = B = 0$, and the behavior $e^{-\beta x} \to \infty$ as $x \to -\infty$ implies that we need $C = D = 0$. Thus, (53) is eliminated entirely, and we are left with (50).

COMMENT 3. Thus, this example is a *boundary-value* problem in which the homogeneous solution drops out by virtue of the boundary conditions. Example 3 is an *initial-value* problem in which the homogeneous solution is not zero, but tended to zero as $t \to \infty$, leaving the particular solution as the steady-state solution. ▮

17.3.4. Complex exponential form for Fourier series.

Using the definitions

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \quad \text{and} \quad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}, \tag{54}$$

it is possible to re-express the Fourier series formula (5) in terms of complex exponentials, as follows:

$$\boxed{\text{FS}\, f = \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/\ell},} \tag{55a}$$

where

$$\boxed{c_n = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx.} \tag{55b}$$

Although the c_n's and exponentials in (55a) are complex, the series does have a real-valued sum.

Notice that the usual definition

$$\sum_{n=1}^{\infty} A_n \equiv \lim_{N\to\infty} \sum_{n=1}^{N} A_n$$

does not apply to (55a) because the lower limit is infinite as well. From our derivation of (55), below, we will see that the appropriate meaning of the series in (55a) is

$$\sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/\ell} \equiv \lim_{N\to\infty} \sum_{n=-N}^{N} c_n\, e^{in\pi x/\ell}. \tag{56}$$

Let us proceed:

$$\begin{aligned}\text{FS}\, f &= a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\\ &= \lim_{N\to\infty} \sum_{n=0}^{N}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\end{aligned}$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} \left(a_n \frac{e^{in\pi x/\ell} + e^{-in\pi x/\ell}}{2} + b_n \frac{e^{in\pi x/\ell} - e^{-in\pi x/\ell}}{2i} \right)$$

$$= \lim_{N\to\infty} \left[\sum_{n=0}^{N} \left(\frac{a_n - ib_n}{2} \right) e^{in\pi x/\ell} + \sum_{n=0}^{N} \left(\frac{a_n + ib_n}{2} \right) e^{-in\pi x/\ell} \right]. \quad (57)$$

Changing n to $-n$ in the second sum, which is permissible because n is merely a dummy summation index,

$$\text{FS}\, f = \lim_{N\to\infty} \left[\sum_{n=0}^{N} \left(\frac{a_n - ib_n}{2} \right) e^{in\pi x/\ell} + \sum_{n=0}^{-N} \left(\frac{a_{-n} + ib_{-n}}{2} \right) e^{in\pi x/\ell} \right]$$

$$= \lim_{N\to\infty} \sum_{n=-N}^{N} c_n \, e^{in\pi x/\ell}$$

$$= \sum_{n=-\infty}^{\infty} c_n \, e^{in\pi x/\ell}. \quad (58)$$

For $n = 0$,

$$c_0 = a_0 = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, dx, \quad (59)$$

for $n > 0$,

$$c_n = \frac{a_n - ib_n}{2} = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x) \left(\cos \frac{n\pi x}{\ell} - i \sin \frac{n\pi x}{\ell} \right) dx$$

$$= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx, \quad (60)$$

and for $n < 0$,

$$c_n = \frac{a_{-n} + ib_{-n}}{2} = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x) \left[\cos\left(-\frac{n\pi x}{\ell} \right) + i \sin\left(-\frac{n\pi x}{\ell} \right) \right] dx$$

$$= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx. \quad (61)$$

All three cases, (59)–(61), agree with (55b), as was to be shown.

The complex form (55) is sometimes favored, especially for electrical engineering applications. An advantage of (55) over (5) is that (55) contains only the exponentials and one set of coefficients (the c_n's), whereas (5) contains both cosines and sines and two sets of coefficients (the a_n's and b_n's).

EXAMPLE 5. Let us return to the square wave of Example 1 and work out its complex Fourier series. With $\ell = \pi$, and f displayed in Fig. 1,

$$c_n = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx = \frac{1}{2\pi} \left[\int_{-\pi}^{0} 0\, dx + \int_{0}^{\pi} 4\, e^{-inx}\, dx \right]$$

$$= \frac{2}{\pi}\left(\frac{e^{-inx}}{-in}\right)\Bigg|_{x=0}^{x=\pi} = \begin{cases} 0, & n = \pm 2, \pm 4, \dots \\ -\dfrac{4i}{n\pi}, & n = \pm 1, \pm 3, \dots . \end{cases} \tag{62}$$

For $n = 0$, the quantity following the third equal sign is indeterminate; it is $(1-1)/0 = 0/0$. In that case, in place of $\int 4e^{-inx}\,dx = 4e^{-inx}/(-in)$ we should use $\int 4e^{0}\,dx = 4x$, which gives $c_0 = 2$. Thus,

$$f(x) = 2 + \sum_{n=-\infty\,(n\text{odd})}^{\infty} \left(-\frac{4i}{n\pi}\right) e^{inx}, \tag{63}$$

and it is not difficult to verify that (63) is equivalent to (11). ▮

Closure. The Fourier series of a 2ℓ-periodic function f, given by (5), converges to $f(x)$ at every point x at which f is continuous, if f and f' are piecewise continuous on $[-\ell, \ell]$. These conditions are sufficient, not necessary, and are met by virtually all functions of applied interest. If f is discontinuous at x, then the series converges to the mean value $[f(x+) + f(x-)]/2$. If the latter does not equal the defined value of $f(x)$, then a discrepancy exists at that point. However, such pointwise discrepancies will be inconsequential in applications.

Beyond the question of convergence, the *speed* of convergence depends on how well-behaved the function is and is of great practical importance. Surely it matters to us whether we need to add 100, 000 terms to achieve a desired accuracy, or whether three or four terms will suffice.

Specifically, if a periodic function is C^N, then we can expect its Fourier coefficients to tend to zero at least as fast as $1/n^{N+2}$. In the loaded beam example (Example 4), for instance, the solution $u(x)$ is C^3, so its Fourier coefficients decay like $1/n^5$. Because of that rapid decay, it suffices to retain only the first two or three terms of the series.

The Fourier coefficients of a given periodic function often contain such quantities as $\cos n\pi$, $\sin n\pi$, $\cos{(n\pi/2)}$, $\sin{(n\pi/2)}$, $\cos{(3n\pi/4)}$, and so on. Of these, $cosn\pi = (-1)^n$ and $\sin n\pi = 0$, but the others are not so readily expressed in algebraic form, so we will leave them intact, and we suggest that you do the same – in working out the exercises.

EXERCISES 17.3

1. Are these functions piecewise continuous on $[0, \pi]$? Explain.

(a) $\sin^2 x$ (b) $\tan x$

(c) $\sin \dfrac{1}{x}$ (d) $\cos \dfrac{1}{x}$

(e) e^{-x} (f) $1/(x-1)$

(g) $1/x$ (h) $\sqrt{x}$

(i) $\begin{cases} 100, & x \neq 2 \\ 50, & x = 2 \end{cases}$ (j) $\begin{cases} x, & 0 \leq x \leq 2 \\ 3 - x, & 2 < x \leq \pi \end{cases}$

2. (a) Derive (24a). HINT: $\cos A \cos B = \frac{1}{2}[\cos{(A+B)} + \cos{(A-B)}]$

(b) Derive (24b). HINT: $\sin A \sin B = \frac{1}{2}[\cos{(A-B)} - \cos{(A+B)}]$

(c) Derive (24c).

3. If both left- and right-hand derivatives of f exist at x_0 and are equal, does that imply that f is differentiable at x_0? Explain.

4. Work out the Fourier series of f, given over one period as follows. At which values of x, if any, does the series fail to converge to $f(x)$? To what values does it converge at those points?

(a) x on $(-\pi, \pi]$
(b) $|x|$ on $(-\pi, \pi]$
(c) $|x|$ on $(-2\pi, 2\pi]$
(d) 50 on $(0, 2]$
(e) 50 on $(0, 2)$, 100 at $x = 2$
(f) 50 on $-8 < x < -2$, 0 on $-2 \leq x \leq 4$
(g) $|\sin x|$ (for all x)
(h) $|\cos x|$ (for all x)
(i) $\sin x$ on $0 \leq x < \pi$, 0 on $\pi \leq x < 2\pi$
(j) $20 + 3 \sin 4x$ (for all x)
(k) x on $0 < x < 1$, 1 on $1 \leq x \leq 2$
(l) $\cos^2 x$ on $0 < x \leq \pi$ HINT: $\cos^2 x = (1 + \cos 2x)/2$
(m) $\sin^2 x$ on $0 < x \leq \pi$
(n) e^{-x} on $0 \leq x < 2$
(o) 100 on $0 \leq x < 1$, 50 on $1 \leq x < 2$, 0 on $2 \leq x < 3$

5. Work out the Fourier series of the 2π-periodic function f defined on $-\pi < x \leq \pi$ as follows, using computer software (such as the *Maple* int command) to evaluate the integrals. HINT: For parts (c) and (d) you will need to use l'Hôpital's rule to simplify the result.

(a) x^2 (b) x^3 (c) $\cos^2 x$ (d) $\sin^2 x$

6. Obtain a computer plot of the partial sums of the Fourier series (19) of the periodic function shown in Fig. 6, for

(a) $n = 2$ (b) $n = 5$ (c) $n = 10$
(d) $n = 20$ (e) $n = 30$ (f) $n = 50$

7. (a) Use (11) to show that

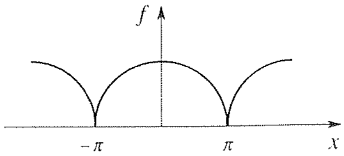
$$\frac{\pi}{4} = \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin \frac{n\pi}{2} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots. \tag{7.1}$$

(b) If you were to use (7.1) to compute π, how many terms would be needed to obtain accuracy to six significant figures? Explain.

8. Let f be the periodic function shown in the figure, each segment of which is a semicircle of radius π. Show that its Fourier series is

$$f(x) = \frac{\pi^2}{4} + \frac{\pi^2}{2} \sum_{n=1}^{\infty} [J_0(n\pi) + J_2(n\pi)] \cos nx, \tag{8.1}$$

where the J_n's denote *Bessel functions of the first kind.*



HINT: You may use any of these formulas:

$$\cos(a \sin\theta) = J_0 + 2J_2 \cos 2\theta + 2J_4 \cos 4\theta + \cdots,$$
$$\sin(a \sin\theta) = 2J_1 \sin\theta + 2J_3 \sin 3\theta + 2J_5 \sin 5\theta + \cdots,$$
$$\cos(a \cos\theta) = J_0 - 2J_2 \cos 2\theta + 2J_4 \cos 4\theta - \cdots,$$
$$\sin(a \cos\theta) = 2J_1 \cos\theta - 2J_3 \cos 3\theta + 2J_5 \cos 5\theta - \cdots,$$

where the J_n's are shorthand for $J_n(a)$. NOTE: Observe that Theorem 17.3.1 does not apply in this case since f' is not piecewise continuous on $[-\pi, \pi]$, but it follows from a stronger version of that theorem that the series in (8.1) does converge to $f(x)$ for all x.

9. If the Fourier coefficients of a periodic function $f(x)$ are a_n $(n = 0, 1, 2, \ldots)$ and b_n $(n = 1, 2, \ldots)$, what are the Fourier coefficients A_n, B_n, say, of the shifted periodic function $f(x - c)$?

10. We have been interpreting the period 2ℓ, in equations (5), to be the *fundamental* period. However, surely (5) should yield the same result if we use twice the fundamental period, and so on. The purpose of this exercise is not to prove that claim but merely to illustrate its truth through a concrete example. Specifically, use (5) to determine the Fourier series of the square wave shown in Fig. 1, using $2\ell = 4\pi$ (rather than 2π), and show that you obtain exactly the same final result as was given in (11).

11. (*Polynomials*) It is a useful fact that if $p(x)$ is an even polynomial (i.e., containing only even-integer powers of x) on $(-\ell, \ell)$, then the b_n's are zero and

$$a_n = \frac{2\ell}{n^2\pi^2} (-1)^n \left[p'(\ell) - \frac{\ell^2}{n^2\pi^2} p'''(\ell) + \frac{\ell^4}{n^4\pi^4} p^{(v)}(\ell) - \cdots \right] \tag{11.1}$$

for $n \geq 1$; and if $p(x)$ is an odd polynomial on $(-\ell, \ell)$, then the a_n's are zero and

$$b_n = -\frac{2}{n\pi} (-1)^n \left[p(\ell) - \frac{\ell^2}{n^2\pi^2} p''(\ell) + \frac{\ell^4}{n^4\pi^4} p^{(iv)}(\ell) - \cdots \right] \tag{11.2}$$

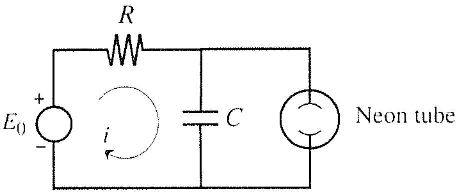
for $n \geq 1$.

(a) Derive (11.1) through the p''' term.
(b) Derive (11.2) through the p'' term.

12. Use (11.1) and/or (11.2) in Exercise 11 to obtain the Fourier series of the given periodic function. NOTE: (11.1) does not hold for $n = 0$, so you need to compute a_0 in the usual way.

(a) $f(x) = x$ on $(-3, 3)$
(b) $f(x) = x^2$ on $(-3, 3)$
(c) $f(x) = 6 + 2x - x^3$ on $(-2, 2)$
(d) $f(x) = x^2(x^2 - 8)$ on $(-2, 2)$
(e) $f(x) = x(x^2 - 3)$ on $(-1, 1)$
(f) $f(x) = 2 + x - x^2$ on $(-3, 3)$
(g) $f(x) = x + x^2$ on $(-1, 1)$

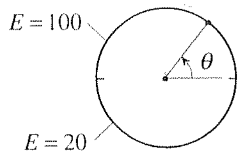
13. (*Relaxation oscillator*) There exist a great many periodic occurrences, called **relaxation oscillations**, which are characterized by a slow buildup and a rapid discharge. Examples include epidemics, economic crises, the sound generated by the bowing of a violin string, shivering from the cold, menstruation, the beating of the heart, and the discharge of a capacitor through a neon tube. Here we consider the latter (see the following figure). The resistance R, capacitance C, and voltage E_0 are constants. If $Q(t)$ is the charge on the capacitor, then $i(t) = dQ/dt$ is the current in the E_0, R, C loop, and Kirchhoff's law gives



$$R\frac{dQ}{dt} + \frac{1}{C}Q = E_0. \tag{13.1}$$

(a) Solve (13.1) subject to the initial condition $Q(0) = 0$.
(b) When the voltage Q/C on the capacitor, and hence on the neon tube, reaches a certain level, say $E_0/2$, the neon tube fires, and the charge Q on the capacitor drops to zero. The firing is so rapid compared to the "buildup time" that we may assume it to be instantaneous. From your solution to part (a), show that the neon tube fires at $t = RC \ln 2$.
(c) The cycle then repeats indefinitely. Sketch and label the graph of the periodic function $Q(t)$, and work out its Fourier series.

14. The voltage $E(\theta)$ is maintained at 100 volts on the top edge of the disk and at 20 volts on the bottom edge. Expand $E(\theta)$ in a Fourier series. HINT: Sketch the graph of $E(\theta)$ on $-\infty < \theta < \infty$.



15. (*Rms current*) If a steady electric current i flows through a resistor of resistance R, the power delivered (i.e., the rate of doing work) is equal to i^2R. In many applications i is not a constant, but a periodic function of the time t. In such cases one defines the *average power* as

$$\text{average power} = \frac{1}{T}\int_{-T/2}^{T/2} i^2(t)R\,dt, \tag{15.1}$$

where T is the fundamental period of $i(t)$. Expressing $i(t)$ as a Fourier series,

$$i(t) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T}\right), \tag{15.2}$$

show that the

$$\text{average power} = \left[a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2)\right]R, \tag{15.3}$$

and that the steady current that is equivalent to $i(t)$, in that it will deliver the same power, is

$$I_{\text{rms}} = \sqrt{a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2)}. \tag{15.4}$$

I_{rms} is known as the **root-mean square (rms) current.**

16. (*Complex form*) Work out the complex exponential form of the Fourier series (55), for the given periodic function, defined over one period as follows.

(a) $f(x) = 50$ on $|x| \le 1$, and 0 on $1 < |x| \le 2$
(b) $f(x) = e^x$ on $0 \le x < 2$
(c) $f(x) = e^{-x}$ on $-3 \le x < 3$
(d) $f(x) = 6 \sin x$ (for all x)
(e) $f(x) = 4 - 5\cos 2x$ (for all x)
(f) $f(x) = -100$ on $-5 \le x < 0$, and 100 on $0 \le x < 5$
(g) $f(x) = 20 + 5x$ on $-2 \le x < 2$

17. (*Example 3*) (a) In Example 3 we expand $F(t)$ in a Fourier series. Could we have used a Taylor series, profitably, instead?
(b) It is proposed that we solve (30) in a piecewise manner. That is, since $F(t)$ is piecewise constant we might do well to consider the t intervals $0 < t < a$, $a < t < 2\pi - a$, $2\pi - a < t < 2\pi + a$, and so on. Solving $mx'' + cx' + kx$

$= \frac{1}{2a}$ on $0 < t < a$, subject to initial conditions $x(0) = x_0$ and $x'(0) = x_0'$, say, we can then use the final values $x(a)$ and $x'(a)$ from the first interval as the initial conditions for the next interval, $a < t < 2\pi - a$, on which $mx'' + cx' + kx = 0$, and so on. You don't need to carry out this method; rather, we merely ask you to assess whether it is a good idea for determining the steady-state response.
(c) Would the method used in Example 3 work if (30) were modified to

$$mx'' + cx' + \alpha x + \beta x^3 = F(t)?$$

Explain.
(d) If we let $a \to 0$ in Example 3, then (see Fig. 8) $F(t)$ becomes a series of delta functions (unit impulses), at $t = 0, 2\pi, 4\pi, \ldots$. (Actually, the one at $t = 0$ will be *half* a delta function.) Find the Fourier series representation of the steady-state response.
(e) Same as part (d), but find a closed form expression of the steady-state response $x(t)$ over one period, and sketch its graph.

18. Consider an undamped, driven mechanical oscillator, governed by the equation

$$mx'' + kx = F(t).$$

Solve for the steady-state response $x(t)$, if $F(t)$ is periodic and defined (over one period) as follows. Let $m = k = 1$.

(a) $F(t) = 100$ on $0 \leq t < 2$, and 0 on $2 \leq t < 4$
(b) $F(t) = 30$ on $0 \leq t < 2$, and -30 on $2 \leq t < 4$
(c) $F(t) = 5t$ on $0 \leq t < 1$, and $10 - 5t$ on $1 \leq t < 2$
(d) $F(t) = 5t$ on $0 \leq t < 1$, and $5t - 10$ on $1 \leq t < 2$
(e) $F(t) = 20$ on $0 \leq t < 2$, and 10 on $2 \leq t < 4$, and 0 on $4 \leq t < 6$
(f) $F(t) = 10t$ on $0 \leq t < 3$, and 0 on $3 \leq t < 4$

17.4 Half- and Quarter-Range Expansions

It often happens in applications, especially when we solve partial differential equations by the method of separation of variables (Chapters 18–20), that we need to expand a given function f in a Fourier series, where f is defined only on a finite interval such as the function f whose graph is shown in Fig. 1, and which is defined only on $0 < x < L$.* But in that case f is not periodic, so how can we expand it in a Fourier series?

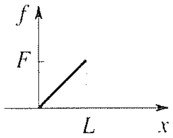


Figure 1. f on finite interval.

The idea is to extend the domain of definition of f to $-\infty < x < \infty$ and to define an "extended function," say f_{ext}, so that f_{ext} is periodic, with $f_{\text{ext}}(x) = f(x)$ on the original interval $0 < x < L$. There are an infinite number of such extensions, two of which are shown in Fig. 2. Each of these is periodic, the first with period $2L$ and the second with period L. Their Fourier series are different, but each of them converges to the original function f on the original interval $0 < x < L$.

How, then, do we know which extension to use? We shall see that the choice will be dictated by the context, so let us not worry about that right now. We will always need to choose from among four extensions which are known as half- and quarter-range cosine and sine extensions and which are based on symmetry or antisymmetry about the endpoints $x = 0$ and $x = L$. These are shown, for the representative function f of Fig. 1, in Fig. 3. For instance, in Fig. 3a f_{ext} is symmetric about $x = 0$ and also about $x = L$, hence the two S's below the x axis.

*We use the open interval notation throughout this section since the values of $f(x)$ at the endpoints will not affect the Fourier coefficients, as we learned in the foregoing section.

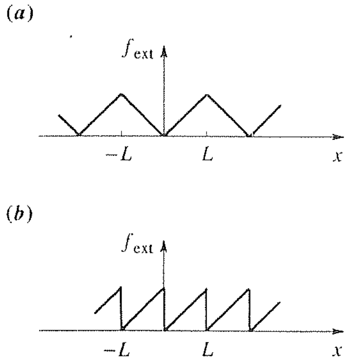


Figure 2. Possible extensions.

Because of its symmetry about $x = 0$, f_{ext} is an even function, and its Fourier series will contain only cosines, no sines. Further, its period is $2L$, so L is half the period. Thus, it is customary to designate this case as the "half-range cosine extension," which we denote in this text by the letters HRC. In Fig. 3b, f_{ext} is antisymmetric about $x = 0$ and $x = L$, the period is $2L$, and we have the half-range sine extension, denoted by HRS. In Fig. 3c f_{ext} is symmetric about $x = 0$ and antisymmetric about $x = L$, the period is $4L$ (so L is only a quarter of the period), and we have the quarter-range cosine extension, denoted by QRS. Similarly for the quarter-range sine extension shown in Fig. 3d and denoted by QRS.

Let us derive the Fourier series for these cases. For the half-range cosine case the period is $2L$, so $\ell = L$ (where we carry over the ℓ notation from Section 17.3) and

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f_{\text{ext}}(x)\,dx = \frac{1}{L}\int_{0}^{L} f(x)\,dx, \tag{1a}$$

$$a_n = \frac{1}{L}\int_{-L}^{L} f_{\text{ext}}(x)\cos\frac{n\pi x}{L}\,dx = \frac{2}{L}\int_{0}^{L} f(x)\cos\frac{n\pi x}{L}\,dx, \tag{1b}$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f_{\text{ext}}(x)\sin\frac{n\pi x}{L}\,dx = 0. \tag{1c}$$

The last step in (1a) follows because $f_{\text{ext}}(x)$ is symmetric about $x = 0$. Similarly, the symmetry of $f_{\text{ext}}(x)\cos(n\pi x/L)$ about $x = 0$ gives (1b), and the antisymmetry of $f_{\text{ext}}(x)\sin(n\pi x/L)$ about $x = 0$ gives (1c). Thus, we can write

$$f_{\text{ext}}(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}, \qquad (-\infty < x < \infty) \tag{2}$$

with a_0 and a_n given by (1a) and (1b).

Understand that the extension was only an artifice, to make possible a Fourier representation of the original function f on the original interval $0 < x < L$. With the expansion in hand, we can now limit our attention to the $0 < x < L$ interval and write

$$\boxed{\begin{aligned} f(x) &= a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}, \qquad (0 < x < L)\\ a_0 &= \frac{1}{L}\int_0^L f(x)\,dx, \qquad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx. \end{aligned}} \tag{3}$$

We call (3) the **half-range cosine expansion** of f. Observe that the final formulas in (3) contain no artifacts of the extension, only f defined on $0 < x < L$.

By a similar process we obtain the **half-range sine expansion**

$$\boxed{\begin{aligned} f(x) &= \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}, \qquad (0 < x < L)\\ b_n &= \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx, \end{aligned}} \tag{4}$$

the **quarter-range cosine expansion**

$$\boxed{\begin{aligned} f(x) &= \sum_{n=1,3,\dots}^{\infty} a_n \cos\frac{n\pi x}{2L}, \qquad (0<x<L) \\ a_n &= \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{2L}\,dx, \end{aligned}} \tag{5}$$

and the **quarter-range sine expansion**

$$\boxed{\begin{aligned} f(x) &= \sum_{n=1,3,\dots}^{\infty} b_n \sin\frac{n\pi x}{2L}, \qquad (0<x<L) \\ b_n &= \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{2L}\,dx. \end{aligned}} \tag{6}$$

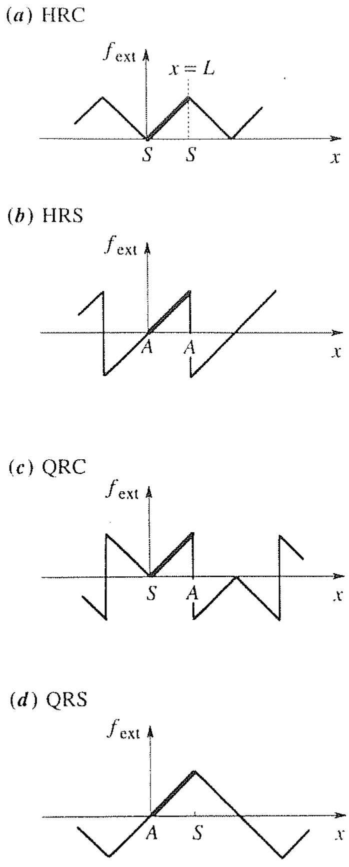


Figure 3. Half- and quarter-range extensions.

We have already derived (3). Let us now derive (6) and leave (4) and (5) for the exercises. We see from Fig. 3d that the period is $4L$, so $\ell = 2L$. Thus

$$f_{\text{ext}}(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{2L} + b_n\sin\frac{n\pi x}{2L}\right). \tag{7}$$

The antisymmetry of f_{ext} about $x=0$ implies that $a_0 = 0$ and $a_n = 0$ for each $n = 1, 2, \dots$. Next,

$$b_n = \frac{1}{2L}\int_{-2L}^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx = \frac{2}{2L}\int_0^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx \tag{8}$$

because both $f_{\text{ext}}(x)$ and $\sin(n\pi x/2L)$ are symmetric about $x=0$, so their product is symmetric about $x=0$.* Now consider the symmetry or antisymmetry about $x = L$. We see from Fig. 3d that $f_{\text{ext}}(x)$ is symmetric about that point. Further, it is easily verified that $\sin(n\pi x/2L)$ is symmetric about that point if n is odd, and antisymmetric about that point if n is even. Thus, the integrand in the second integral in (8) is symmetric about $x = L$ (which is the midpoint of the integration interval) if n is odd, and antisymmetric about that point if n is even. Hence (8) gives

$$\begin{aligned} b_n &= \frac{1}{L}\int_0^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx \\ &= \begin{cases} 0 & , \quad n \text{ even} \\ \dfrac{2}{L}\displaystyle\int_0^L f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx\ , & n \text{ odd.} \end{cases} \end{aligned} \tag{9}$$

*Understand the difference between symmetric/antisymmetric and even/odd. A graph can be symmetric about *any* x point. If, in particular, the graph of f is symmetric/antisymmetric about the *origin*, then f is even/odd, respectively.

Finally, we can drop the subscripted "ext" in the final integral in (9) because the interval of integration is $0 < x < L$, over which interval $f_{\text{ext}}(x) = f(x)$. Putting these expressions for a_0, a_n, and b_n into (7) gives (6).

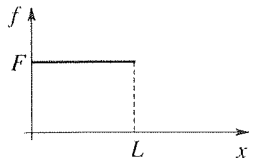


Figure 4. f in Example 1.

EXAMPLE 1. To illustrate, let us expand the function f, displayed in Fig. 4, in half- and quarter-range cosine and sine series.

HRC: From (3),

$$\begin{aligned} a_0 &= \frac{1}{L}\int_0^L F\,dx = \frac{FL}{L} = L, \\ a_n &= \frac{2}{L}\int_0^L F\cos\frac{n\pi x}{L}\,dx = \frac{2F}{n\pi}\sin\frac{n\pi x}{L}\Big|_{x=0}^{x=L} = 0, \end{aligned} \tag{10}$$

so the HRC expansion of f is simply

$$f(x) = F, \tag{11}$$

which results from the extension shown in Fig. 5a.

HRS: From (4),

$$b_n = \frac{2}{L}\int_0^L F\sin\frac{n\pi x}{L}\,dx = -\frac{2F}{n\pi}(\cos n\pi - 1), \tag{12}$$

so the HRS expansion of f is

$$f(x) = \frac{2F}{\pi}\sum_{n=1}^{\infty}\frac{1 - cosn\pi}{n}\sin\frac{n\pi x}{L}, \tag{13}$$

which results from the extension shown in Fig. 5b.

QRC: From (5),

$$a_n = \frac{2}{L}\int_0^L F\cos\frac{n\pi x}{2L}\,dx = \frac{4F}{n\pi}\sin\frac{n\pi}{2}, \tag{14}$$

so the QRC expansion of f is

$$f(x) = \frac{4F}{\pi}\sum_{n=1,3,\dots}^{\infty}\frac{sin(n\pi/2)}{n}\cos\frac{n\pi x}{2L}, \tag{15}$$

which results from the extension shown in Fig. 5c.

QRS: From (6),

$$b_n = \frac{2}{L}\int_0^L F\sin\frac{n\pi x}{2L}\,dx = \frac{4F}{n\pi}(1 - \cos\frac{n\pi}{2}) = \frac{4F}{n\pi} \tag{16}$$

because $\cos\frac{n\pi}{2} = 0$ if n is odd, and n is indeed odd in (6). Thus, the QRS expansion of f is

$$f(x) = \frac{4F}{\pi}\sum_{n=1,3,\dots}^{\infty}\frac{1}{n}\sin\frac{n\pi x}{2L}, \tag{17}$$

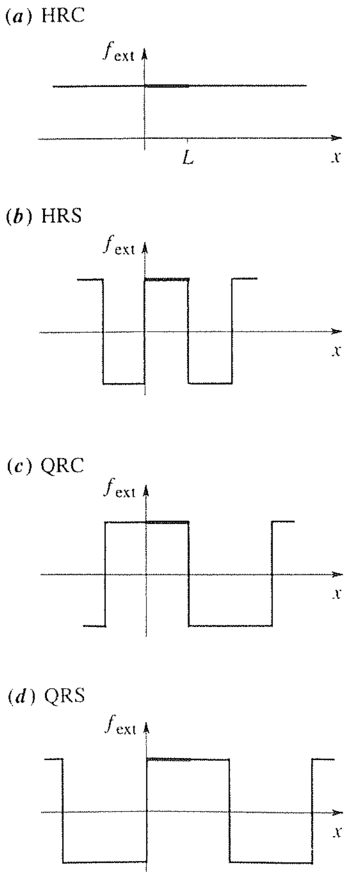


Figure 5. Half- and quarter-range extensions of f.

which results from the extension shown in Fig. 5d.

The series (11), (13), (15), and (17) converge to the functions shown in Fig. 5a,b,c,d, respectively, but on the interval $0 < x < L$ they all converge to the function f defined in Fig. 4.

COMMENT. Observe that n runs from 1 to infinity continuously in the half-range formulas (3) and (4), but only through the odd values $1, 3, \ldots$ in the quarter-range formulas (5) and (6). In the present example, the sum in the HRS expansion (13) runs through $n = 1, 3, \ldots$ because $1 - \cos n\pi = 1 - (-1)^n = 0$ for even n's, but in general the sums in the half-range expansions run through $n = 1, 2, 3, \ldots$. ▮

Closure. Functions defined only on a finite interval, say $0 < x < L$, can be periodically extended in an infinite number of ways. Of these, only four possible extensions will prove to be of interest, the half- and quarter-range extensions, which lead to the half- and quarter-range cosine and sine series (3)–(6). Application of these results occur repeatedly in Chapters 18–20 when we solve partial differential equations by the method of separation of variables.

EXERCISES 17.4

1. We derived the HRC and QRS formulas, (3) and (6), respectively. In similar fashion, derive the

(a) HRS formulas (4) (b) QRC formulas (5)

2. For the given function, prepare labeled sketches of the HRC, HRS, QRC, and QRS extensions, and derive the corresponding expansions.

(a) $f(x) = 25$ on $0 < x < 1$, and 0 on $1 < x < 2$

(b) $f(x) = 5x$ on $0 < x < 4$

(c) $f(x) = \sin x$ on $0 < x < \pi$

(d) $f(x) = \sin x$ on $0 < x < \pi/2$

(e) $f(x) = 1 - x$ on $0 < x < 1$

(f) $f(x) = 0$ on $0 < x < 4$, and 50 on $4 < x < 5$

(g) $f(x) = \sin x$ on $0 < x < \pi$, and 0 on $\pi < x < 2\pi$

(h) $f(x) = 2 + x$ on $0 < x < 3$

17.5 Manipulation of Fourier Series (Optional)

In Examples 3 and 4 of Section 17.4 we differentiated Fourier series termwise and equated coefficients of trigonometric series on the left- and right-hand sides of an equation. Our approach is formal (i.e., we do not rigorously justify the steps) since we did not yet have theorems in place with which to justify those steps. This section is intended to provide the necessary theorems.

Fundamental is the concept of uniform convergence.* Consider a series

$$\sum_{n=1}^{\infty} a_n(x), \tag{1}$$

and let $s_n(x)$ be its nth partial sum

$$s_n(x) = a_1(x) + \cdots + a_n(x). \tag{2}$$

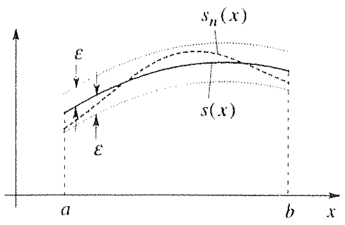


Figure 1. Uniform convergence.

We say that (1) **converges uniformly** to $s(x)$ on an interval $a \leq x \leq b$ if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds an $N(\epsilon)$, independent of x, such that $|s_n(x) - s(x)| < \epsilon$ for all $n > N$ and for all x in the interval $a \leq x \leq b$.† The situation is illustrated in Fig. 1. Supposing that $s(x)$ is as shown, choose an arbitrarily small $\epsilon > 0$ and draw an "ϵ band" about $s(x)$, the band between $s(x) - \epsilon$ and $s(x) + \epsilon$. If (1) converges uniformly, then (no matter how small we choose ϵ to be) there must be some integer N such that the graph of $s_n(x)$ lies entirely within the ϵ band for all n's greater than N.

EXAMPLE 1. Let (1) be the Fourier series

$$f(x) = 2 + \frac{8}{\pi} \sum_{n=1,,3,\ldots}^{\infty} \frac{1}{n} \sin nx \tag{3}$$

of the square wave, one period of which is shown in Fig. 2. We know that that series converges to the square wave shown in the figure: 0 for $-\pi < x < 0$, 2 at $x = 0$ (denoted by the heavy dot), and 4 for $0 < x < \pi$). Is that convergence uniform? The answer depends upon the interval under consideration. Over $1 \leq x \leq 2$, for instance, the answer is yes, which we state without proof; for any $\epsilon > 0$, no matter how small, we can force the graph of $s_n(x)$ to fall within the ϵ band by taking n sufficiently large.

However, over any interval containing a jump discontinuity of f, such as $-1 \leq x \leq 1$, the answer is no because the partial sums are continuous, so the graph of $s_n(x)$ must inevitably break out of the ϵ bands (at P and Q) in order to pass through the heavy dot. In addition, there are breakouts due to the Gibbs phenomenon, the spike-like overshoots and undershoots near the jump discontinuities. ▮

Of course we cannot rely on pictures, we need an analytical test for uniformity

*The concept of uniform convergence is due to *Karl Weierstrass* (1841) and *G. G. Stokes* (1847). Generally speaking, it is probably true that major theorems attract the most acclaim in mathematics but, in truth, the fundamental definitions – such as the definitions of the *limit* of a function, the *derivative* of a function, the *uniform convergence* of a series, and so on, are of comparable importance. It is easy to invent definitions, but not all definitions are fruitful. Consider, for instance, the way all of the differential and integral calculus, as well as other branches of analysis, rest upon the shoulders of the limit concept.

†Contrast this definition with the definition of convergence of (1) at x: (1) converges to $s(x)$ at x if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds an $N(\epsilon)$ such that $|s_n(x) - s(x)| < \epsilon$ for all $n > N$.

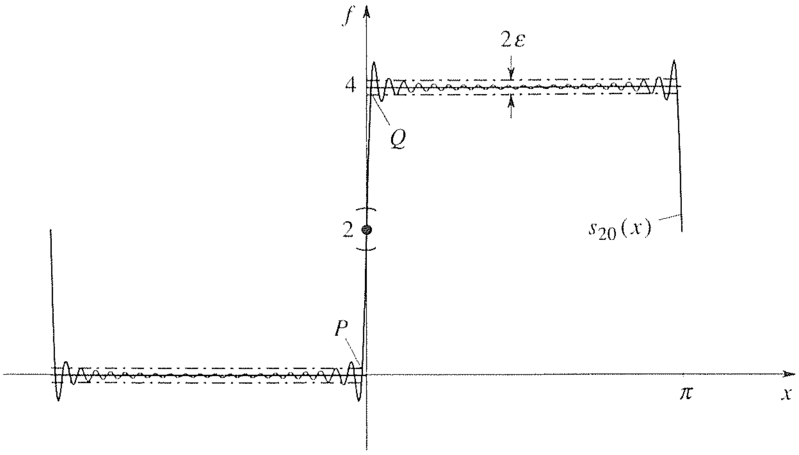


Figure 2. Square wave Fourier series.

of convergence. A useful sufficient condition for uniform convergence is as follows.

THEOREM 17.5.1 *Weierstrass M-Test*
If $\sum_{n=1}^{\infty} M_n$ is a convergent series of positive constants and $|a_n(x)| \leq M_n$ on an x interval I, then $\sum_{n=1}^{\infty} a_n(x)$ is uniformly (and absolutely) convergent on I.

EXAMPLE 2. The series

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} \tag{4}$$

converges uniformly on $-\infty < x < \infty$, according to the Weierstrass M-test, because $|(\cos nx)/n^2| \leq 1/n^2$ on $-\infty < x < \infty$, and $\sum_{n=1}^{\infty} 1/n^2$ is convergent; specifically, it is a convergent p-series.* That is, M_n is $1/n^2$ in this case. ▮

We can now state a useful theorem on the termwise differentiation of an infinite series.

THEOREM 17.5.2 *Termwise Differentiation of Series*
Let $\sum_{n=1}^{\infty} a_n(x)$ converge on an x interval I. Then

$$\boxed{\frac{d}{dx} \sum_{n=1}^{\infty} a_n(x) = \sum_{n=1}^{\infty} \frac{d}{dx} a_n(x)} \tag{5}$$

*Recall from the calculus that the **p-series** $\sum_{n=1}^{\infty} 1/n^p$ converges if $p > 1$ and diverges if $p \leq 1$. If $p = 1$ it becomes the well known (and divergent) **harmonic series**.

if the series on the right converges uniformly on I.*

That is, we can interchange the order of the two limit operations (the infinite series is the limit of the sequence of partial sums, and the derivative is the limit of a difference quotient) if the $a_n(x)$'s are well enough behaved, specifically, if $\sum a_n'(x)$ converges uniformly on I. That condition is stated as sufficient, not necessary. To put (5) into perspective, it might be helpful to recall the analogous result for interchanging the order of differentiation and integration: if a and b are constants, then

$$\frac{d}{dx}\int_a^b f(t,x)\,dt = \int_a^b \frac{\partial f}{\partial x}(t,x)\,dt \tag{6}$$

if f is sufficiently well behaved, namely, if f and $\partial f/\partial x$ are continuous on the relevant rectangle in the x, t plane (see the Leibniz rule, Theorem 13.8.1).

EXAMPLE 3. The series

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^3} \tag{7}$$

can be differentiated termwise on $-\infty < x < \infty$. That is,

$$\frac{d}{dx}\sum_{n=1}^{\infty} \frac{\sin nx}{n^3} = \sum_{n=1}^{\infty} \frac{d}{dx}\left(\frac{\sin nx}{n^3}\right) = \sum_{n=1}^{\infty} \frac{\cos nx}{n^2}, \tag{8}$$

because the series on the right converges uniformly for all x, as shown in Example 2. ▮

Besides the termwise differentiations employed in the representative Examples 3 and 4 of Section 17.3, we also equated the corresponding coefficients of trigonometric series on the left- and right-hand sides of an equation. Justification of that step is provided by the following theorem.

THEOREM 17.5.3 *Uniqueness of Trigonometric Series*
If

$$a_0+\sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right) = A_0+\sum_{n=1}^{\infty}\left(A_n \cos\frac{n\pi x}{\ell} + B_n \sin\frac{n\pi x}{\ell}\right), \tag{9}$$

where the trigonometric series on the left- and right-hand sides converge to the

*For proof of this theorem and additional discussion of this issue, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison–Wesley, 1957), Chap.13.

same sum for all x, then $a_0 = A_0$, $a_n = A_n$, and $b_n = B_n$ for each n.*

Notice the wording "trigonometric series" for the theorem. Recall that a trigonometric series is a series of the form given by the left- and right-hand sides of (9). Why don't we say Fourier series instead? Because not every convergent trigonometric series is a Fourier series. For example,

$$\sum_{n=1}^{\infty} \frac{1}{\ln(n+1)} \sin nx \tag{10}$$

is a trigonometric series, and it can be shown (by a Dirichlet test) to converge for all x, yet it is not a Fourier series.* That is, there does not exist a 2π-periodic function f such that

$$\int_{-\pi}^{\pi} f(x) \cos nx \, dx = 0, \qquad (n = 0, 1, 2, \dots) \tag{11a}$$

$$\int_{-\pi}^{\pi} f(x) \sin nx \, dx = \frac{1}{\ln(n+1)}. \qquad (n = 1, 2, \dots) \tag{11b}$$

Thus, every Fourier series is a trigonometric series, but not every convergent trigonometric series is a Fourier series. This result is in interesting contrast with the result, encountered later in this text (Theorem 24.2.8), that every convergent power series is the Taylor series of its sum function. However, if a trigonometric series, with period 2ℓ, converges *uniformly* on $[-\ell, \ell]$ (or, equivalently, on $-\infty < x < \infty$), then it is the Fourier series of its sum function.†

Let us illustrate these results with a practical application that is similar to Examples 3 and 4 in Section 17.3.

EXAMPLE 4. Find a particular solution to the differential equation

$$x'' + 0.5x = f(t), \qquad (0 < t < \infty) \tag{12}$$

where the forcing function $f(t)$ is the 2π-periodic function shown in Fig. 3. The Fourier of f is

$$f(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin(2n-1)t, \tag{13}$$

and if we seek $x(t)$ in the form

$$x(t) = \sum_{n=1}^{\infty} b_n \sin(2n-1)t \tag{14}$$

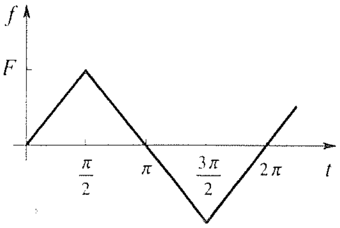


Figure 3. $f(t)$ in Example 4.

*Proof of this sophisticated theorem can be found in E. C. Titchmarsh, *The Theory of Functions*, 2nd ed. (London: Oxford University Press, 1939), pp. 427–432.

*See I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, 2nd ed. (New York: McGraw Hill, 1966), p. 86, Exer.3.

†See G. P. Tolstov, *Fourier Series* (Englewood Cliffs, NJ: Prentice–Hall, 1962), pp. 14–15.

we find (Exercise 1), formally, that

$$x(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)^2[(2n-1)^2 - 0.5]} \sin(2n-1)t. \tag{15}$$

We can make the process rigorous either by justifying each of the steps in the derivation of (15a), using Theorems 17.5.1–17.5.3, or we can proceed formally to (15) and then rigorously verify that the latter does satisfy the given differential equation (12). Let us do the latter.

According to Theorem 17.5.2 we can differentiate (15) termwise and obtain

$$x'(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)[(2n-1)^2 - 0.5]} \cos(2n-1)t \tag{16}$$

because the latter series is uniformly convergent on $0 < t < \infty$. That is,

$$\left| \frac{(-1)^n}{(2n-1)[(2n-1)^2 - 0.5]} \cos(2n-1)t \right| \leq \frac{1}{(2n-1)[(2n-1)^2 - 0.5]} = M_n, \tag{17}$$

and $\sum_{n=1}^{\infty} M_n$ is convergent because $M_n \sim 1/8n^3$ as $n \to \infty$, where $\sum_{n=1}^{\infty}(1/8n^3) = \frac{1}{8}\sum_{n=1}^{\infty} 1/n^3$ is $\frac{1}{8}$ times a convergent p-series, with $p = 3$.* Likewise, we can differentiate (16) termwise and obtain

$$x''(t) = -\frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)^2 - 0.5} \sin(2n-1)t \tag{18}$$

because the latter series is uniformly convergent on $0 < t < \infty$. That is,

$$\left| \frac{(-1)^n}{(2n-1)^2 - 0.5} \sin(2n-1)t \right| \leq \frac{(-1)^n}{(2n-1)^2 - 0.5} = M_n, \tag{19}$$

where $\sum_{n=1}^{\infty} M_n$ is convergent because $M_n \sim 1/4n^2$ as $n \to \infty$.

Finally, putting (18), (15), and (13) into (12) and adding the two series on the left-hand side termwise† is found to produce an identity, so the verification of (15) is complete.

COMMENT. Observe that each time we differentiate (15) we weaken the convergence because we pull out a $2n-1$ from the sine or cosine. Since$f$ was C^0 (i.e., continuous), the terms in (13) were of order $O(1/n^2)$ and, consequently, the terms in (15) were $O(1/n^4)$, as $n \to \infty$. Thus, the terms in $x'(t)$ were $O(1/n^3)$ and those in $x''(t)$ were $O(1/n^2)$, the n^2 being sufficient to render the series uniformly convergent so as to justify the termwise differentiation of (16). If $f(t)$ had been even better behaved such as C^1 or C^2, then so much the better, but if it had been *dis*continuous such as a square wave, then our efforts at justification would have failed because then the terms in the $x''(t)$ series would die out like

*Recall from the calculus that if $\sum a_n$ is a series of positive constants and $a_n \sim Kb_n$as $n \to \infty$, for some positive constant K, then the series $\sum a_n$ and $\sum b_n$ both converge or both diverge.

†Recall from the calculus that if $\sum a_n$ and $\sum b_n$ are convergent, then $\sum a_n + \sum b_n = \sum(a_n + b_n)$, and the latter is convergent as well.

$O(1/n)$, which would not be fast enough to ensure uniform convergence by the Weierstrass M-test (since the series $\sum 1/n$ is divergent). ▮

Actually, the idea of finding solutions to differential equations that are in the form of infinite series is not new to us since we have already studied *power series* solutions in Chapter 4. Indeed, the technical issues that arose there are the same: the termwise differentiation of a power series, the termwise addition of two series, and the fact that if two power series are equal, then their corresponding coefficients must be equal. These matters are handled by parts (a), (b), (c), respectively, of Theorem 4.2.3. There, we saw that power series are very easy to work with, especially in that a power series may be differentiated (or integrated) any number of times (within its interval of convergence). Of course, that fact is not surprising since a convergent power series is the Taylor series of its sum function, and for a function to have a Taylor series it must be extremely well-behaved, for instance, infinitely differentiable (C^∞). However, in applications like Example 4, above, power series are of no help because $f(t)$ is merely continuous; it is not even once differentiable because of the kinks in its graph at $t = \pi/2, 3\pi/2, 5\pi/2, \ldots$. But since f is periodic we are able to use Fourier series.

The fact that the Fourier series (15) of $x(t)$ can be differentiated termwise only twice is not surprising since $x(t)$ is not infinitely differentiable, it is only C^2 (since the input f is C^0 and x is obtained from f, in effect, by two integrations).

Though we have not used termwise integration of Fourier series we would be remiss if we did not include the following theorem.

THEOREM 17.5.4 *Termwise Integration of Fourier Series*
If a Fourier series is integrated termwise between any finite limits, the resulting series converges to the integral of the periodic function corresponding to the original series.

EXAMPLE 5. Let us integrate the square wave f shown in Fig. 4, say, from 0 to any point x. From Example 1 of Section 17.3 we have, for the Fourier series of f,

$$f(x) = 2 + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin nx. \tag{20}$$

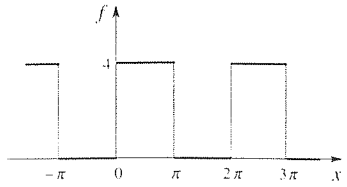


Figure 4. Square wave f.

Thus,

$$\begin{aligned} \int_0^x f(\xi)\,d\xi &= \int_0^x \left[2 + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin n\xi \right] d\xi \\ &= \int_0^x 2\,d\xi + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \int_0^x \sin n\xi \, d\xi \end{aligned}$$

$$= 2x + \frac{8}{\pi} \sum_{n=1,3,\dots}^{\infty} \frac{1-\cos nx}{n^2}, \tag{21}$$

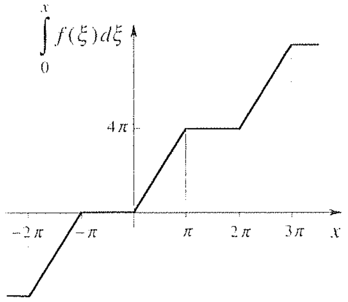


Figure 5. Sum of the integrated series.

where the second equality follows from Theorem 17.5.4. Further, the final right-hand side, in (21) must, according to the theorem, converge to $\int_0^x f(\xi)\, d\xi$, the graph of which can be inferred by inspection of Fig. 4 and is drawn in Fig. 5.

COMMENT. The series (21) is not quite a Fourier series because of the $2x$ term, which arises because the average value of f [corresponding to the 2 in (20)] is nonzero. ▮

Closure. This section is about technical matters related to the manipulation of Fourier series, especially regarding their termwise differentiation. The latter is important because functions represented by Fourier series are often not very well behaved; for instance, they may be C^1, C^0, or even discontinuous. And since differentiation makes bad behavior even worse,* one is well advised to approach the differentiation of Fourier series with caution. Theorem 17.5.2 tells us that termwise differentiation is permissible if the resulting series is uniformly convergent, although that condition is stated as sufficient, not necessary. In turn, Theorem 17.5.1 gives us a simple and well known test for uniformity of convergence, the Weierstrass M-test.

Whereas each termwise differentiation of a Fourier series pulls a factor of n out of the sine and cosine terms, thus retarding the speed of convergence, termwise integration introduces factors of $1/n$ and therefore enhances the convergence. In fact, Theorem 17.5.4 tells us that *every* Fourier series can be safely integrated termwise.

EXERCISES 17.5

1. Assuming the form (14), derive the solution (15).

2. Show that the given series is uniformly convergent on the given interval.

(a) $\sum_1^\infty e^{-nx} \sin nx$ on $2 < x < 5$

(b) $\sum_1^\infty \dfrac{\sin 2nx}{n^2 - 2n + 2}$ on $-\infty < x < \infty$

(c) $\sum_0^\infty \dfrac{e^{-nx}}{n^2+5}$ on $0 < x < \infty$

(d) $\sum_3^\infty \dfrac{1}{n^2+x^2}$ on $-\infty < x < \infty$

(e) $\sum_1^\infty \ln\left(1 + \dfrac{x^3}{n^3}\right)$ on every finite interval

2. Show that $\sum_1^\infty n^{-x}$ converges uniformly on $x_0 \le x < \infty$, for every $x_0 > 1$.

3. Determine an x interval on which the series is uniformly convergent. NOTE: You may need to use one of the standard convergence tests, from the calculus.

(a) $\sum_1^\infty \dfrac{1}{1+x^n}$

(b) $\sum_1^\infty n(\cos x)^n$

(c) $\sum_1^\infty n^5(\sin x)^n$

(d) $\sum_6^\infty (5x)^n$

(e) $\sum_1^\infty \dfrac{x^n}{n^3}$

(f) $\sum_1^\infty (x^2 + 2x - 1)^n$

(g) $\sum_3^\infty (x^2 - 2x)^{3n}$

4. Differentiate the given series termwise, and verify the validity of that step if x is in the given interval (which is not necessarily the broadest one possible).

(a) $\sum_1^\infty \dfrac{\sin 2nx}{n^4}$, $\quad -\infty < x < \infty$

*See Comment 1 in Example 4, Section 17.3.

(b) $4 + \sum_1^\infty \dfrac{\cos n\pi x}{n^3+1}, \quad -\infty < x < \infty$

(c) $\sum_1^\infty (4x)^n, \quad -0.2 \le x \le 0.1$

(d) $\sum_1^\infty \dfrac{x^n}{n^3}, \quad -1 \le x \le 1$

(e) $\sum_1^\infty e^{-n} \sin nx, \quad -\infty < x < \infty$

5. Derive a particular solution to the forced oscillator equation

$$x'' + x' + x = f(t), \qquad (0 < t < \infty) \qquad (5.1)$$

where $f(t)$ is periodic and defined over one period as follows, and then rigorously verify that your solution does indeed satisfy (5.1).

(a) $f(t) = \sin t$ on $0 \le t \le \pi$
(b) $f(t) = |\cos t|$ on $0 \le t \le \pi$
(c) $f(t) = \sin t$ on $0 \le t \le \pi$ and 0 on $\pi \le t \le 2\pi$
(d) $f(t) = |5 - 5t|$ on $0 \le t \le 2$
(e) $f(t) = 0$ on $0 \le t \le \pi/2$ and $\cos t$ on $\pi/2 \le t \le 3\pi/2$

6. Evaluate $\int_0^1 f(x)\,dx$ to three significant figures, and verify the validity of the termwise integration.

(a) $f(x) = \dfrac{4}{\pi} \displaystyle\sum_{n=1}^{\infty} \dfrac{\cos(2n-1)x}{(2n-1)^2}$

(b) $f(x) = \dfrac{2}{\pi} - \dfrac{4}{\pi} \displaystyle\sum_{n=1}^{\infty} \dfrac{\cos 2nx}{4n^2 - 1}$

17.6 Vector Space Approach

An elegant and more modern approach to Fourier series is available within the vector space context. Vector space is the subject of Section 9.6, which section you may wish to review before continuing.

Specifically, consider the **function space** $C_p[a, b]$ of all real-valued piecewise-continuous functions defined on $[a, b]$, that is, on $a \le x \le b$. First, let us verify that $C_p[a, b]$ is indeed a vector space. Let $\mathbf{f} = f(x)$ and $\mathbf{g} = g(x)$ be any two functions* in $C_p[a, b]$, and let α be any (real) scalar. We define the sum $\mathbf{f} + \mathbf{g}$ and the scalar multiple $\alpha\mathbf{f}$ by

$$\mathbf{f} + \mathbf{g} \equiv f(x) + g(x), \qquad \alpha\mathbf{f} \equiv \alpha f(x), \qquad (1)$$

respectively. Observe that if f and g are piecewise continuous on $[a, b]$ then so is $f + g$, so that $C_p[a, b]$ is closed under vector addition. Similarly, if f is piecewise continuous on $[a, b]$ then so is αf, so that $C_p[a, b]$ is closed under scalar multiplication. Furthermore, we define the zero vector $\mathbf{0}$ as the function which is identically zero, so that $\mathbf{f} + \mathbf{0} = f(x) + 0 = f(x) = \mathbf{f}$. And we define the negative inverse of $\mathbf{f} = f(x)$ as $-\mathbf{f} \equiv -f(x)$, in which case we have $\mathbf{f} + (-\mathbf{f}) = f(x) + [-f(x)] = 0 = \mathbf{0}$.

With these definitions of vector addition, scalar multiplication, the zero vector, and the negative inverse, we can see that all of the requirements listed in Definition 9.6.1 are satisfied, so that $C_p[a, b]$ is indeed a vector space.

Next, we wish to introduce an inner product for $C_p[a, b]$ and we choose the

*By "$\mathbf{f} = f(x)$" we mean that the vector $\mathbf{f}$ is the function whose values are $f(x)$. Notation becomes tricky here since there are now three quantities to be distinguished: the function f considered as a mapping, the values $f(x)$ of that function, and the vector $\mathbf{f}$.

inner product of f and g as[†]

$$\langle \mathbf{f}, \mathbf{g} \rangle \equiv \int_a^b f(x)g(x)\,dx, \tag{2}$$

which is introduced in Example 4 of Section 9.6. Recall that f and g are **orthogonal** if $\langle \mathbf{f}, \mathbf{g} \rangle = 0$, that the **norm** $\|\mathbf{f}\|$ of f is defined as $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle}$, and that f is said to be **normalized** (i.e., scaled so as to have unit norm) if $\|\mathbf{f}\| = 1$.

Finally, recall from our study of best approximation, in Section 9.10, the following important result: If f is any vector in a normed inner product vector space $\mathcal{S}$ with natural norm $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle}$, and $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ is an ON (orthonormal) set in $\mathcal{S}$, then the **best approximation** of f within span $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ is given by the **orthogonal projection** of f onto span $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$, namely, by

$$\mathbf{f} \approx \langle \mathbf{f}, \hat{\mathbf{e}}_1 \rangle \hat{\mathbf{e}}_1 + \cdots + \langle \mathbf{f}, \hat{\mathbf{e}}_N \rangle \hat{\mathbf{e}}_N = \sum_{n=1}^{N} \langle \mathbf{f}, \hat{\mathbf{e}}_n \rangle \hat{\mathbf{e}}_n. \tag{3}$$

To apply these results to Fourier series, let $\mathcal{S}$ be $C_p[a, b]$, with the inner product and norm defined above, let $a = -\ell$ and $b = \ell$, and consider the vectors

$$\mathbf{e}_1 = 1,\ \ \mathbf{e}_2 = \cos\frac{\pi x}{\ell},\ \ \mathbf{e}_3 = \sin\frac{\pi x}{\ell},\ \ \cdots,\ \ \mathbf{e}_{2k} = \cos\frac{k\pi x}{\ell},\ \ \mathbf{e}_{2k+1} = \sin\frac{k\pi x}{\ell} \tag{4}$$

in $C_p[-\ell, \ell]$. The set $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_N\}$ is orthogonal by virtue of the inner product definition (2) and the integrals (24a,b,c) in Section 17.3. For instance,

$$\langle \mathbf{e}_2, \mathbf{e}_3 \rangle = \int_{-\ell}^{\ell} \cos\frac{\pi x}{\ell} \sin\frac{\pi x}{\ell}\,dx = 0. \tag{5}$$

Furthermore,

$$\begin{aligned} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle &= \int_{-\ell}^{\ell} (1)(1)\,dx = 2\ell, \\ \langle \mathbf{e}_2, \mathbf{e}_2 \rangle &= \int_{-\ell}^{\ell} \cos^2\frac{\pi x}{\ell}\,dx = \ell, \\ &\vdots \\ \langle \mathbf{e}_{2k+1}, \mathbf{e}_{2k+1} \rangle &= \int_{-\ell}^{\ell} \sin^2\frac{k\pi x}{\ell}\,dx = \ell, \end{aligned} \tag{6}$$

[†] If we were considering a complex function space (i.e., where the functions are complex-valued and the scalars are complex), then we would use the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle \equiv \int_a^b f(x)\overline{g}(x)\,dx$$

instead of (2), in accordance with the discussion in Section 12.2, where $\overline{g}(x)$ is the complex conjugate of $g(x)$. For example, if $g(x) = 6 + 5i$, then $\overline{g}(x) = 6 - 5i$.

so that the normalized $\mathbf{e}_n$'s are

$$\begin{aligned}\hat{\mathbf{e}}_1 &= \frac{1}{\|\mathbf{e}_1\|}\,\mathbf{e}_1 = \frac{1}{\sqrt{\langle \mathbf{e}_1, \mathbf{e}_1\rangle}}\,\mathbf{e}_1 = \frac{1}{\sqrt{2\ell}},\\ \hat{\mathbf{e}}_2 &= \frac{1}{\|\mathbf{e}_2\|}\,\mathbf{e}_2 = \frac{1}{\sqrt{\langle \mathbf{e}_2, \mathbf{e}_2\rangle}}\,\mathbf{e}_2 = \frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell},\\ &\vdots\\ \hat{\mathbf{e}}_{2k+1} &= \frac{1}{\sqrt{\ell}}\sin\frac{k\pi x}{\ell}.\end{aligned} \tag{7}$$

Since the set $\{\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_{2k+1}\}$ is ON, it is LI (linearly independent), and since k (and hence $2k+1$) is arbitrarily large, it follows that we have an arbitrarily large number of LI vectors in $C_p[-\ell, \ell]$, so the latter is *infinite dimensional*.

Returning to (3), we are approximating a given $\mathbf{f} = f(x)$, in $C_p[-\ell, \ell]$, in the form

$$\begin{aligned}f(x) \approx c_1\frac{1}{\sqrt{2\ell}} + c_2\frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell} + c_3\frac{1}{\sqrt{\ell}}\sin\frac{\pi x}{\ell} + \cdots\\ + c_{2k}\frac{1}{\sqrt{\ell}}\cos\frac{k\pi x}{\ell} + c_{2k+1}\frac{1}{\sqrt{\ell}}\sin\frac{k\pi x}{\ell},\end{aligned} \tag{8}$$

where

$$\begin{aligned}c_1 &= \langle \mathbf{f}, \hat{\mathbf{e}}_1\rangle = \int_{-\ell}^{\ell} f(x)\frac{1}{\sqrt{2\ell}}\,dx = \frac{1}{\sqrt{2\ell}}\int_{-\ell}^{\ell} f(x)\,dx\\ c_2 &= \langle \mathbf{f}, \hat{\mathbf{e}}_2\rangle = \int_{-\ell}^{\ell} f(x)\frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell}\,dx = \frac{1}{\sqrt{\ell}}\int_{-\ell}^{\ell} f(x)\cos\frac{\pi x}{\ell}\,dx\\ &\vdots\\ c_{2k+1} &= \langle \mathbf{f}, \hat{\mathbf{e}}_{2k+1}\rangle = \frac{1}{\sqrt{\ell}}\int_{-\ell}^{\ell} f(x)\sin\frac{k\pi x}{\ell}\,dx.\end{aligned} \tag{9}$$

Equivalently, we can write

$$\begin{aligned}f(x) &\approx a_0 + a_1\cos\frac{\pi x}{\ell} + b_1\sin\frac{\pi x}{\ell} + \cdots + a_k\cos\frac{k\pi x}{\ell} + b_k\sin\frac{k\pi x}{\ell}\\ &= a_0 + \sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell} + b_n\sin\frac{n\pi x}{\ell}\right),\end{aligned} \tag{10}$$

where

$$a_0 = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx, \tag{11a}$$

$$a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, \tag{11b}$$

$$b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx. \tag{11c}$$

Let us review the situation. The right-hand side of (10) is the orthogonal projection of $\mathbf{f}$ on the $(2k+1)$-dimensional subspace of $C_p[-\ell, \ell]$, which is the span of the ON vectors

$$\frac{1}{\sqrt{2\ell}}, \quad \frac{1}{\sqrt{\ell}} \cos\frac{\pi x}{\ell}, \quad \frac{1}{\sqrt{\ell}} \sin\frac{\pi x}{\ell}, \dots, \frac{1}{\sqrt{\ell}} \cos\frac{k\pi x}{\ell}, \quad \frac{1}{\sqrt{\ell}} \sin\frac{k\pi x}{\ell}. \tag{12}$$

With the coefficients in (10) equal to the by-now-familiar Fourier coefficients, as given in (11), (10) is the best possible approximation to $f(x)$ of the form (8). That is, it is the best approximation in the sense of minimizing the norm of the error vector

$$\left\| \mathbf{f} - \left(c_1 \frac{1}{\sqrt{2\ell}} + c_2 \frac{1}{\sqrt{\ell}} \cos\frac{\pi x}{\ell} + \cdots + c_{2k+1} \frac{1}{\sqrt{\ell}} \sin\frac{k\pi x}{\ell} \right) \right\|^2$$
$$= \int_{-\ell}^{\ell} \left[f(x) - \left(c_1 \frac{1}{\sqrt{2\ell}} + c_2 \frac{1}{\sqrt{\ell}} \cos\frac{\pi x}{\ell} + \cdots + c_{2k+1} \frac{1}{\sqrt{\ell}} \sin\frac{k\pi x}{\ell} \right) \right]^2 dx. \tag{13}$$

The integral in (13) is known as the **squared error**. Thus (10) is the best approximation in the sense of minimizing the squared error and is called the **least-square approximation** of $f(x)$ with respect to the ON set (4).

We can well expect the squared error to diminish as we increase k, but the key question is: Does it tend to zero as $k \to \infty$? It does.

THEOREM 17.6.1 *Vector Convergence*
If $f(x)$ is piecewise continuous on $[-\ell, \ell]$ and $a_0, a_1, b_1, a_2, b_2, \dots$ are the Fourier coefficients defined by (11), then

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell} \right) \tag{14}$$

holds in the sense of vector (least-square) convergence, namely,

$$\lim_{k\to\infty} \int_{-\ell}^{\ell} \left[f(x) - a_0 - \sum_{n=1}^{k} \left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell} \right) \right]^2 dx = 0. \tag{15}$$

Proof of this sophisticated theorem is well beyond our present scope.

Note that the meaning of (14) here is different from its meaning in preceding sections. Specifically, in preceding sections equation (14) held in the pointwise sense, namely, that at a specific single fixed value of x we have

$$\lim_{k\to\infty} \left[a_0 + \sum_{n=1}^{k} \left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell} \right) \right] = f(x) \tag{16}$$

or, equivalently,

$$\boxed{\lim_{k\to\infty}\left[f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]=0.} \tag{17}$$

In contrast, in the present section (14) is understood as a vector equation, which holds in the sense that

$$\boxed{\lim_{k\to\infty}\int_{-\ell}^{\ell}\left[f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]^2 dx=0.} \tag{18}$$

The truth of (17) does not imply the truth of (18), nor does the truth of (18) imply the truth of (17); they are logically independent statements (Exercise 4).

To better appreciate the distinction, observe that the pointwise convergence addressed in Theorem 17.3.1 is convergence in a *local* sense, at a specific point x, whereas the vector convergence addressed in Theorem 17.6.1 is convergence in a *global* sense. That is, the squared error *integrated over the entire interval* tends to zero as $k\to\infty$.

We do not wish to imply that one form of convergence is inherently better or more correct than the other; they are simply different.

Closure. In effect, Theorem 17.6.1 tells us that the infinite set of orthogonal vectors $\{1,\cos\frac{\pi x}{\ell},\sin\frac{\pi x}{\ell},\cos\frac{2\pi x}{\ell},\sin\frac{2\pi x}{\ell},\ldots\}$ comprises an orthogonal *basis* for $C_p[-\ell,\ell]$. Couldn't we say that that set is a basis because $C_p[-\ell,\ell]$ is infinite-dimensional and there is an infinite number of vectors in the set? No. For suppose we remove one (or more) of the vectors from the set. Then we still have an infinite set of orthogonal vectors, but they are not a basis. For instance, if we remove the $\cos(\pi x/\ell)$ vector and take $f(x)$ to be $6\cos(\pi x/\ell)$, say, then (14) would be

$$6\cos\frac{\pi x}{\ell}=0+0+0+\cdots,$$

which is surely incorrect.

EXERCISES 17.6

1. Corresponding to the approximation (10), let the error vector be

$$\mathbf{E}=f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right). \tag{1.1}$$

(a) Show that

$$\|\mathbf{E}\|^2=\int_{-\ell}^{\ell}[f(x)]^2\,dx-\ell\left[2a_0^2+\sum_{n=1}^{k}(a_n^2+b_n^2)\right]. \tag{1.2}$$

(b) Deduce, from (1.2), the **Bessel inequality**

$$\boxed{2a_0^2 + \sum_{n=1}^{k}(a_n^2 + b_n^2) \le \frac{1}{\ell}\int_{-\ell}^{\ell}[f(x)]^2\,dx,} \qquad (1.3)$$

which holds for each $k = 1, 2, \dots$. NOTE: Since $\|\mathbf{E}\| \to 0$ as $k \to \infty$, according to Theorem 17.6.1, we obtain

$$\boxed{2a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) = \frac{1}{\ell}\int_{-\ell}^{\ell}[f(x)]^2\,dx,} \qquad (1.4)$$

which is known as the **Parseval equality**, and which is an infinite-dimensional function space version of the familiar Pythagorean theorem. We can draw an interesting and useful conclusion from (1.4). The integral on the right converges because f has been assumed piecewise continuous on $[-\ell, \ell]$. Thus, the series on the left converges and has nonnegative terms. For such a series we know that its nth term must tend to zero as $n \to \infty$. Hence, it follows that $a_n \to 0$ and $b_n \to 0$ as $n \to \infty$.

2. Use equation (1.2), above, to compute $\|\mathbf{E}\|$, for $k = 1, 2, \dots, 8$, where the 2π-periodic function f is defined on $-\pi < x < \pi$ as follows.

(a) $f(x) = |x|$ (b) $f(x) = x$
(c) $f(x) = \cos^2 x$ (d) $f(x) = x^2$
(e) $f(x) = |\sin x|$ (f) $f(x) = |\cos x|$

(g) $f(x) = \begin{cases} 0, & -\pi < x < 0 \\ 1, & 0 < x < \pi \end{cases}$

(h) $f(x) = \begin{cases} 0, & -\pi < x < 0 \\ x, & 0 < x < \pi \end{cases}$

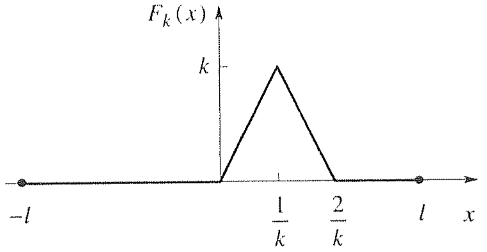
(i) $f(x) = \begin{cases} 100, & -\pi < x < \pi/2 \\ 50, & \pi/2 < x < \pi \end{cases}$

(j) $f(x) = \begin{cases} \sin x, & -\pi < x < 0 \\ 0, & 0 < x < \pi \end{cases}$

3. Let f and g both be members of $C_p[-\ell, \ell]$, and $f(x) = g(x)$ on $[-\ell, \ell]$ except at a finite number of points, at which $f(x) \neq g(x)$. Show that $\mathbf{f} = \mathbf{g}$ in spite of these pointwise differences; i.e., show that $\|\mathbf{f} - \mathbf{g}\| = 0$, so that within the vector space framework $\mathbf{f}$ and $\mathbf{g}$ are indistinguishable.

4. (*Vector convergence and pointwise convergence*) Below Theorem 17.6.1, we emphasized that vector and pointwise convergence are independent, neither one implies the other. The purpose of this exercise is to illustrate that claim through a simple example. Consider, in place of the messy $[\ ]^2$ integrand in (18), the sequence $F_k(x)$ defined in parts (a), (b), and (c).

(a) First, consider the sequence displayed below.



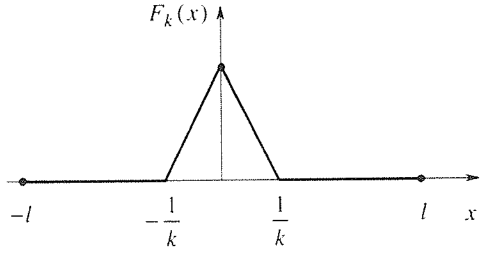
As k increases, the "mountain" becomes taller and narrower. Show that

$$\lim_{k\to\infty} F_k(x) = 0 \qquad (4.1a)$$

on $[-\ell, \ell]$, whereas

$$\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 1 \neq 0. \qquad (4.1b)$$

(b) With $F_k(x)$ as shown below, instead, show that



$$\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 0, \qquad (4.1c)$$

whereas

$$\lim_{k\to\infty} F_k(x) = \begin{cases} 0, & x \neq 0 \\ 1, & x = 0 \end{cases} \neq 0. \qquad (4.1d)$$

(c) Let $F_k(x)$ be the same as in part (b), except that the height of the "mountain" is $1/k$ instead of 1. Show that in this case *both* limits are zero. CONCLUSION: The statements $\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 0$ and $\lim_{k\to\infty} F_k(x) = 0$ on $[-\ell, \ell]$ are *independent*; the truth of one does not imply the truth of the other.

5. Beginning with the expression

$$\|\mathbf{E}\|^2 = \int_{-\ell}^{\ell}\left[f(x) - a_0 - \sum_{n=1}^{k}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\right]^2 dx$$

for the square of the norm of the error vector $\mathbf{E}$ associated with the approximation (10), seek the optional choice of a_0, a_n, b_n by setting $\partial \|\mathbf{E}\|^2 / \partial a_0 = 0$, $\partial \|\mathbf{E}\|^2 / \partial a_n = 0$, and $\partial \|\mathbf{E}\|^2 / \partial b_n = 0$. Show that that step produces the expressions for a_0, a_n, b_n given in (11).

17.7 The Sturm–Liouville Theory

17.7.1. Sturm–Liouville problem. In Section 17.6 we found that the sines and cosines present in Fourier series constitute an orthogonal basis for the relevant infinite-dimensional function space. Where do such bases come from? Are there others as well? In this section we discover that such bases arise as the eigenfunctions of Sturm–Liouville eigenvalue problems, just as real symmetric $n \times n$ matrices provide us with sets of eigenvectors that are orthogonal bases for n-space. Thus, Fourier series will be found to be just one (extremely important) example within a broader Sturm–Liouville theory.*

By a **Sturm–Liouville problem** we mean a linear homogeneous second-order differential equation

$$\boxed{[p(x)y']' + q(x)y + \lambda w(x)y = 0, \qquad (a < x < b)} \tag{1a}$$

with homogeneous boundary conditions of the form

$$\boxed{\begin{aligned} \alpha y(a) + \beta y'(a) &= 0, \\ \gamma y(b) + \delta y'(b) &= 0, \end{aligned}} \tag{1b}$$

where a, b are finite, where p, p', q, w are continuous on $[a, b]$, and where $p(x) > 0$ and $w(x) > 0$ on $[a, b]$. These conditions, as well as the precise form of (1a) and (1b), are important, and should be carefully noted. Further, α and β are not both zero, γ and δ are not both zero, and $a, b, p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ are all real.

Also critical is that (1) is a *boundary-value* problem, not an initial-value problem, because the conditions (1b) are imposed at both ends of the interval. If, in place of (1b), we imposed homogeneous initial conditions $y(a) = 0$ and $y(b) = 0$, then (1) would admit the trivial solution $y(x) = 0$, and that solution would be unique (Theorem 3.3.1). However, we saw in Section 3.3.2 that boundary-value problems can admit no solution, a unique solution, or a nonunique solution. Thus, even though the Sturm–Liouville boundary-value problem (1) surely

*The theory presented in this section was discovered by *Charles Sturm* (1803–1855) and *Joseph Liouville* (1809–1882), and published in 1836–1837. Sturm, a professor of mechanics at the Sorbonne, became involved in this work as an outgrowth of his studies of the partial differential equations governing the flow of heat in a bar of nonuniform density. He was joined in this work by his friend Liouville, who was a professor of mathematics at the Collège de France and who is also well known for his work on complex variable theory and on transcendental numbers. Note the spelling Liouville (not Louisville), and the pronunciation lē-ōō′vil (not lōō′ē-vil).

admits the trivial solution, our interest is in finding *nontrivial* solutions. Though $a, b, p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ are all specified, λ is a free parameter. Any value of λ that permits the existence of nontrivial solutions of (1) is called an **eigenvalue** of (1), and the corresponding nontrivial solution is called an **eigenfunction** of (1). Thus, (1) is an **eigenvalue problem**, analogous to the matrix eigenvalue problem studied in Chapter 11.

EXAMPLE 1. Consider the case

$$y'' + \lambda y = 0, \qquad (0 < x < L) \tag{2a}$$

$$y(0) = 0, \quad y(L) = 0. \tag{2b}$$

Comparing (2) with (1) we see that $p(x) = w(x) = 1$, $q(x) = 0$, $a = 0$, $b = L$, $\alpha = \gamma = 1$, and $\beta = \delta = 0$, all of which satisfy the conditions listed below (1).

To see if (2) admits nontrivial solutions, let us not be intimidated by the fact it is an "eigenvalue problem." After all, (2a) is a simple differential equation. Its general solution is evidently*

$$y(x) = A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x. \tag{3}$$

However, if $\lambda = 0$, then (3) reduces to $y(x) = A$, which is not a general solution of (2a). Thus, for the special case $\lambda = 0$ we return to (2a), which becomes $y'' = 0$, and determine the general solution to be $C + Dx$. Thus, in place of (3) we write

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx. & \lambda = 0 \end{cases} \tag{4a,b}$$

Treating the cases $\lambda = 0$ and $\lambda \neq 0$ separately, consider $\lambda = 0$ first. Applying the boundary conditions (2b) to $y(x) = C + Dx$ gives

$$y(0) = 0 = C, \tag{5a}$$
$$y(L) = 0 = C + DL, \tag{5b}$$

so $C = D = 0$. Thus, the only solution corresponding to $\lambda = 0$ is the trivial solution $y(x) = 0$, so $\lambda = 0$ is not an eigenvalue (in this example).

*Note that we don't yet know λ. If $\lambda > 0$, then (3) is the general solution of (2), but does (3) hold if $\lambda < 0$? Shouldn't we have a cosh and sinh in that case? If $\lambda < 0$, (3) becomes,

$$\begin{aligned} y(x) &= A\cos\left(i\sqrt{|\lambda|}\,x\right) + B\sin\left(i\sqrt{|\lambda|}\,x\right) \\ &= A\cosh\left(\sqrt{|\lambda|}\,x\right) + iB\sinh\sqrt{|\lambda|}\,x, \end{aligned}$$

so we do have an arbitrary linear combination of cosh and sinh, as we should. (We could rename iB as C if we don't like the i.) Thus, there is no need to treat the cases $\lambda > 0$ and $\lambda < 0$ separately, although some authors prefer to do that. In fact, (3) holds even if λ is complex, but we will find that for Sturm–Liouville problems λ is always real.

Turning to the case $\lambda \neq 0$, the boundary conditions give

$$y(0) = 0 = A, \tag{6a}$$

$$y(L) = 0 = A\cos\sqrt{\lambda}\,L + B\sin\sqrt{\lambda}\,L. \tag{6b}$$

Since (6a) gives $A = 0$, (6b) gives $B\sin\sqrt{\lambda}\,L = 0$, so either $B = 0$ or $\sin\sqrt{\lambda}\,L = 0$ and B is arbitrary. We rule out the choice $B = 0$ since then we would have $A = B = 0$ and hence the trivial solution $y(x) = 0$. Rather, B is arbitrary and

$$\sin\sqrt{\lambda}\,L = 0. \tag{7}$$

Solving (7), we have $\sqrt{\lambda}L = n\pi$ for $n = 0, \pm1, \pm2, \ldots$. Of these values, discard $n = 0$ because it gives $\lambda = 0$, which case has already been considered. Thus we have the eigenfunctions

$$y(x) = B\sin\frac{n\pi x}{L}, \tag{8}$$

where B is arbitrary and $n = \pm1, \pm2, \ldots$. The negative values of n can be discarded as well since the positive and negative choices do not lead to distinct (i.e., linearly independent) eigenfunctions. For example, $n = +2$ gives $\sin(2\pi x/L)$, and $n = -2$ gives $\sin(-2\pi x/L) = -\sin(2\pi x/L)$, and since the scale factor B is arbitrary it can absorb the minus sign. Let us set $B = 1$, say, for definiteness.

The upshot is that we have the infinite set of eigenvalues and eigenfunctions

$$\lambda_n = \frac{n^2\pi^2}{L^2} \qquad \text{and} \qquad \phi_n(x) = \sin\frac{n\pi x}{L} \tag{9}$$

for $n = 1, 2, \ldots$, where we use the symbol ϕ_n to denote the nth eigenfunction (analogous to the special symbol $\mathbf{e}_n$ that we used in the matrix case).

COMMENT 1. It may be useful to recast the solution of (6) in matrix form because the matrix approach is more convenient in algebraically-more-difficult cases. Re-expressing (6) as

$$\begin{bmatrix} 1 & 0 \\ \cos\sqrt{\lambda}\,L & \sin\sqrt{\lambda}\,L \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{10}$$

we see that we have a nontrivial solution of (10) [and hence of (2)] if and only if the determinant of the coefficient matrix is zero:

$$\begin{vmatrix} 1 & 0 \\ \cos\sqrt{\lambda}\,L & \sin\sqrt{\lambda}\,L \end{vmatrix} = \sin\sqrt{\lambda}\,L = 0, \tag{11}$$

which is the same as (7). Next, solve (11) for the λ's, put them back in (10), and solve (10), say by Gauss elimination, for the nontrivial solutions for A and B. That step gives $A = 0$ and $B =$ arbitrary, as above.

COMMENT 2. We call (11) the **characteristic equation** corresponding to the eigenvalue problem (2). In the $n \times n$ matrix case the characteristic equation is always an nth degree algebraic equation; in the Sturm–Liouville case it is always transcendental (an infinite-degree algebraic equation), with an infinite number of distinct roots.* ▮

Sturm–Liouville problems such as (2) arise throughout Chapters 18–20, when

*To see that (11) is an infinite-degree algebraic equation, replace $\sin\sqrt{\lambda}\,L$ by its Taylor series.

we solve partial differential equations by the method of separation of variables – for instance in connection with unsteady heat conduction in a rod and the lateral motion of a vibrating string. Sometimes, however, they arise directly. For instance, a classical problem in structural mechanics is the determination of the buckling load of a structure. Consider, as a simple structure, the column shown in Fig. 1, of length L, pinned at both ends and subjected to a downward load (i.e., force) P. As we increase P nothing happens – the column remains straight, until a certain value of P is reached, which we call the critical load or *buckling load* and which we denote as P_{cr}. Under that load the column bends (and probably collapses). Clearly, it is important to be able to predict P_{cr}. To do so, we use *Euler beam theory*,[*] which tells us that the lateral deflection $y(x)$ is governed by the boundary-value problem

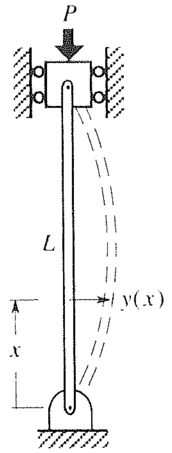


Figure 1. Column buckling.

$$EIy'' + Py = 0, \qquad (0 < x < L) \tag{12a}$$
$$y(0) = 0, \quad y(L) = 0, \tag{12b}$$

where E and I are physical constants of the column: E is Young's modulus of the material and I is the cross-sectional inertia. We see that (12) is the same as (2), with $\lambda = P/EI$. Surely, $y(x) = 0$ satisfies (12), but that solution is of no interest because it does not correspond to buckling. From Example 1 we recall that nontrivial solutions occur for $\lambda = P/EI = \pi^2/L^2, 4\pi^2/L^2, 9\pi^2/L^2, \dots$. The smallest of these, $P/EI = \pi^2/L^2$, gives the buckling load

$$P_{cr} = \frac{\pi^2 EI}{L^2}, \tag{13}$$

and the corresponding eigenfunction $\sin(\pi x/L)$ gives (to within an arbitrary scale factor) the shape of the corresponding buckling mode, which is somewhat as sketched in Fig. 1. The analysis gives the inception of buckling and does not give insight into the dynamical process of collapse. The formula (13) was published first by Euler in 1757.

Since we will be concerned with orthogonal bases in function space, we will need an inner product $\langle f, g\rangle$ between two functions (i.e., vectors) f and g. It will be convenient to use the inner product[†]

$$\boxed{\langle f, g\rangle \equiv \int_a^b f(x)\,g(x)\,w(x)\,dx,} \tag{14}$$

where the weight function $w(x)$ is the $w(x)$ in the Sturm–Liouville equation (1a). If $\langle f, g\rangle = 0$, then f and g are *orthogonal*.

[*] See S. Timoshenko, *Strength of Materials*, Part I (Princeton, NJ: D. Van Nostrand, 1955).

[†] If f and g are complex-valued functions such as e^{ix}, then (14) fails to meet the conditions required of any dot or inner product [(16) in Section 9.6] and should be modified as $\int_a^b f(x)\,\overline{g}(x)\,w(x)\,dx$. We will face up to that detail in the optional Section 17.7.2 but, otherwise, will continue to use (14) because it will turn out that complex-valued functions need not arise. Also, note that whether we write $\langle f, g\rangle$ or $\langle \mathbf{f}, \mathbf{g}\rangle$ doesn't matter, as long as we understand that f and g are here being considered as vectors.

We have the following major theorem regarding the eigenvalues and eigenfunctions of the Sturm–Liouville eigenvalue problem (1), with the restrictions on $p(x), q(x), w(x), a, b, \alpha, \beta, \gamma, \delta$ stated earlier.

THEOREM 17.7.1 *Sturm–Liouville Theorem*
Let λ_n and $\phi_n(x)$ denote any eigenvalue and corresponding eigenfunction of the Sturm–Liouville eigenvalue problem (1), respectively.

(a) The eigenvalues are real.

(b) The eigenvalues are simple. That is, to each eigenvalue there corresponds only one linearly independent eigenfunction. Further, there are an infinite number of eigenvalues, and they can be ordered so that $\lambda_1 < \lambda_2 < \lambda_3 < \cdots$, where $\lambda_n \to \infty$ as $n \to \infty$.

(c) Eigenfunctions corresponding to distinct eigenvalues are orthogonal. That is, if $\lambda_j \neq \lambda_k$, then $\langle \phi_j, \phi_k \rangle = 0$.

(d) Let f and f' be piecewise continuous on $a \leq x \leq b$. If $a_n = \langle f, \phi_n \rangle / \langle \phi_n, \phi_n \rangle$, then the series $\sum_{n=1}^{\infty} a_n \phi_n(x)$ converges to $f(x)$ if f is continuous at x, and to the mean value $[f(x+) + f(x-)]/2$ if f is discontinuous at x, for each point x in the open interval $a < x < b$.

This theorem is analogous to the several individual theorems given in Section 11.3 for $n \times n$ real symmetric matrix eigenvalue problems. Parts of it are proved in the optional Section 17.7.2.

Part (d) says that

$$\boxed{f(x) = \sum_{n=1}^{\infty} \frac{\langle f, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} \phi_n(x)} \tag{15}$$

holds at each point x in $a < x < b$ at which f is continuous. If, at a discontinuity, $f(x)$ does not happen to equal the mean value$[f(x+)+f(x-)]/2$, then the equality in (15) does not hold at that point. To remind us of the possibility of such pointwise discrepancies, some authors write an "ae" above the equal sign, to mean equal "almost everywhere," but we will simply write the eigenfunction expansion of f as we have in (15), without such reminders. Observe that (d) is a pointwise convergence statement. Various other statements are available regarding both pointwise and vector convergence, but (d) will suffice for most practical purposes and for our purposes in this text.

EXAMPLE 2. Consider the results of Example 1 in the light of Theorem 17.7. Since $\lambda_n = n^2\pi^2/L^2$, the eigenvalues are real, there are an infinite number of them, and $\lambda_n \to \infty$ as $n \to \infty$. Further, each eigenvalue is simple because the single eigenfunction

$\phi_n(x) = \sin(n\pi x/L)$ corresponds to each eigenvalue λ_n. Finally, with the weight function $w(x) = 1$ we have

$$\langle \phi_m, \phi_n \rangle = \int_0^L \sin\frac{m\pi x}{L} \sin\frac{n\pi x}{L}\,dx = 0 \tag{16}$$

for $m \neq n$, by virtue of the Euler formula (24b) in Section 17.3. Each of these results is in accord with Theorem 17.7.1.

To illustrate the eigenfunction expansion (15), let $f(x) = x$. Then

$$\langle f, \phi_n \rangle = \int_0^L x \sin\frac{n\pi x}{L}\,dx = (-1)^{n+1}\frac{L^2}{n\pi}, \tag{17a}$$

$$\langle \phi_n, \phi_n \rangle = \int_0^L \sin^2\frac{n\pi x}{L}\,dx = \frac{L}{2}, \tag{17b}$$

so we have

$$f(x) = x = \frac{2L}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin\frac{n\pi x}{L}. \tag{18}$$

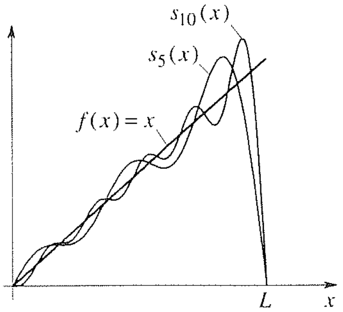


Figure 2. Convergence of (18).

COMMENT 1. Carefully observe that part (d) of the theorem does NOT require f to satisfy the homogeneous Sturm–Liouville boundary conditions, which are $y(0) = 0$ and $y(L) = 0$ in the present example. In fact, $f(L)$ is L, not 0. Nonetheless, we do obtain the convergence that is guaranteed by the theorem, over $0 < x < L$ (actually, over $0 \le x < L$ in this example), as hinted at in Fig. 2, where we compare $f(x) = x$ with the fifth and tenth partial sums, $s_5(x)$ and $s_{10}(x)$.

COMMENT 2. Observe also that, corresponding to the present Sturm–Liouville problem, the eigenfunction expansion (15), namely,

$$f(x) = \sum_{n=1}^{\infty} a_n \sin\frac{n\pi x}{L}, \tag{19a}$$

where

$$a_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx, \tag{19b}$$

is actually the half-range sine expansion of f, studied in Section 17.4. Other choices of the boundary conditions, in place of $y(0) = 0$ and $y(L) = 0$, will result in eigenfunctions that produce the half-range cosine and quarter-range sine and cosine expansions (Exercise 2). ▮

There is a small flaw in our procedure. To solve the characteristic equation (7), we recalled that $\sin x = 0$ has roots at $x = n\pi$ on the real axis. Might there be complex roots as well? It's true that we know in advance that λ is real, but if it is real and negative then the argument of the sine is purely imaginary. Thus, we need to search for roots of $\sin z = 0$ (where $z = x + iy$) not only along the real axis but along the imaginary axis as well. Doing so, observe that $\sin iy = i \sinh y = 0$ is equivalent to $\sinh y = 0$, which admits only the root $y = 0$. Thus, we did not miss any roots, and all is well. The following theorem could have saved us this extra trouble.

THEOREM 17.7.2 *Nonnegative Eigenvalues*
If $q(x) \leq 0$ on $[a, b]$ and $[p(x)\phi_n(x)\phi_n'(x)]|_a^b \leq 0$ for the eigenfunction $\phi_n(x)$, then not only is λ_n real, $\lambda_n \geq 0$.

Applying Theorem 17.7.2 to the problem in Example 1, observe that $q(x) \leq 0$ on $[0, L]$ because $q(x) = 0$. Further, $p(x) = 1$ and the $\phi_n(x)$'s satisfy the boundary conditions (2b), so

$$\begin{aligned} [p(x)\phi_n(x)\phi_n'(x)]\Big|_a^b &= \phi_n(L)\phi_n'(L) - \phi_n(0)\phi_n'(0) \\ &= (0)\phi_n'(L) - (0)\phi_n'(0) = 0. \end{aligned} \tag{20}$$

Hence, not only are the λ_n's real, they are also nonnegative.

EXAMPLE 3. Find the eigenvalues and eigenfunctions for the Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad (0 < x < 1) \tag{21a}$$
$$y(0) - 2y'(0) = 0, \quad y(1) = 0. \tag{21b}$$

We speak of the boundary condition at $x = 0$ as being of **mixed** type because both y and y' are present; that is, both α and β are nonzero in (1b). Such boundary conditions do arise in applications such as unsteady heat conduction, as we shall see in Chapter 18.

As in Example 1, solution of (21a) gives

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx, & \lambda = 0. \end{cases} \tag{22a,b}$$

Imposing the boundary conditions (21b) on the $C + Dx$ solution gives $C - 2D = 0$ and $C + D = 0$, so $C = D = 0$. Hence, $\lambda = 0$ is not an eigenvalue. For the $\lambda \neq 0$ case, the boundary conditions give

$$\begin{bmatrix} 1 & -2\sqrt{\lambda} \\ \cos\sqrt{\lambda} & \sin\sqrt{\lambda} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{23}$$

To obtain nontrivial solutions (i.e., where A and B are not both zero), set the determinant equal to zero, which step gives the characteristic equation $\sin\sqrt{\lambda} + 2\sqrt{\lambda}\cos\sqrt{\lambda} = 0$ or, more conveniently,

$$\tan\sqrt{\lambda} = -2\sqrt{\lambda}. \tag{24}$$

Applying Theorem 17.7.2, with $q(x) = 0$ and $p(x) = 1$,

$$\begin{aligned} (p\phi_n\phi_n')\Big|_0^1 &= \phi_n(1)\phi_n'(1) - \phi_n(0)\phi_n'(0) \\ &= (0) - 2[\phi_n'(0)]^2 \qquad \text{[from (21b)]} \\ &\leq 0, \end{aligned} \tag{25}$$

so the λ_n's are nonnegative. Hence, it suffices to look for solutions of (24) along a real $\sqrt{\lambda}$ axis. Plotting the left- and right-hand sides of (24) in Fig. 3, we can see the roots $\sqrt{\lambda_n}$ fall as indicated. (The intersection at the origin can be discarded because we already determined that $\lambda = 0$ is not an eigenvalue.) We can find any number of roots using computer software such as *Maple* and obtain, for the first few,

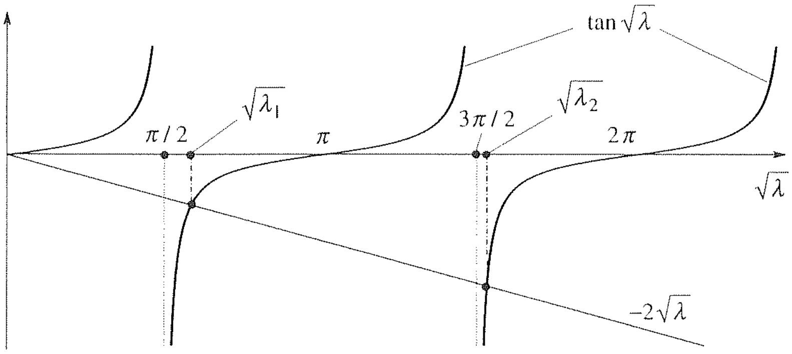


Figure 3. Roots of (24).

$$\sqrt{\lambda_1} = 1.83660, \quad \sqrt{\lambda_2} = 4.81584, \quad \sqrt{\lambda_3} = 7.91705, \ldots, \tag{26}$$

or

$$\lambda_1 = 3.3731, \quad \lambda_2 = 23.1923, \quad \lambda_3 = 62.6797, \ldots, \tag{27}$$

and so on. The graphs in Fig. 3, or even a freehand sketch of them, reveal that $\sqrt{\lambda_n} \sim (2n-1)\pi/2$ as $n \to \infty$.

With the eigenvalues determined, we return to (23) to find the nontrivial solutions for A and B and hence for $y(x)$. With λ satisfying (24), the second row of the coefficient matrix is a multiple of the first, so of the two scalar equations implied by (23) the second can be discarded, leaving the one equation

$$A - 2\sqrt{\lambda_n}\, B = 0 \tag{28}$$

on the two unknowns A and B. Thus, $A = 2\sqrt{\lambda_n}\, B$, where B remains arbitrary, so

$$\begin{aligned} y(x) &= A\cos\sqrt{\lambda_n}\, x + B\sin\sqrt{\lambda_n}\, x \\ &= B(2\sqrt{\lambda_n}\, \cos\sqrt{\lambda_n}\, x + \sin\sqrt{\lambda_n}\, x), \end{aligned} \tag{29}$$

and hence the eigenfunctions are

$$\phi_n(x) = 2\sqrt{\lambda_n}\, \cos\sqrt{\lambda_n}\, x + \sin\sqrt{\lambda_n}\, x, \tag{30}$$

where the λ_n's are given by (27).

COMMENT. Actually, if we extended the graphs in Fig. 3 over $-\infty < \sqrt{\lambda} < 0$ we would find the additional roots $-\sqrt{\lambda_1}$ and $-\sqrt{\lambda_2}$ and so on. These can be discarded just as we discarded the $n = -1, -2, \ldots$ cases in Example 1 because they contribute nothing new.

For instance, the right-hand side of (30) is an odd function of $\sqrt{\lambda_n}$, so changing $\sqrt{\lambda_n}$ to $-\sqrt{\lambda_n}$ merely scales $\phi_n(x)$ by a factor of -1. ▮

EXAMPLE 4. As a final example, consider the problem

$$y'' - 2y' + \lambda y = 0, \qquad (0 < x < \pi) \tag{31a}$$

$$y(0) = 0, \quad y(\pi) = 0. \tag{31b}$$

It appears that (31) may not be a Sturm–Liouville problem at all since the written-out version of (1a) is

$$py'' + p'y' + qy + \lambda wy = 0. \tag{32}$$

That is, *the coefficient of y' needs to be the derivative of the coefficient of y''*. Yet, -2 in (31a) is not the derivative of 1. However, let us multiply (31a) by a yet-to-be-determined function $\sigma(x)$, giving

$$\sigma y'' - 2\sigma y' + \lambda\sigma y = 0 \tag{33}$$

such that the coefficient -2σ of y' is the derivative of the coefficient σ of y'':

$$\sigma' = -2\sigma. \tag{34}$$

Solving, $\sigma(x) = Ce^{-2x}$ and we can take $C = 1$ without loss. Putting that σ into (33) does give the standard Sturm–Liouville form,

$$(e^{-2x}y')' + \lambda e^{-2x}y = 0. \tag{35}$$

Since the factor $\sigma = e^{-2x}$ in (33) is everywhere nonzero, the solution of (31a) and (35) are identical, so the two equations are equivalent.

That step was important for two reasons. First, it establishes the problem as being of Sturm–Liouville type so that we can make use of Theorems 17.7.1 and 17.7.2. Second, it enables us to identify $p(x)$ and $w(x)$: $p(x) = w(x) = e^{-2x}$, so we see that $p(x) > 0$ and $w(x) > 0$ on the closed interval $0 \le x \le \pi$, as required by the theory. And of course we will need to know the weight function $w(x)$ in the inner product if we are to carry out any eigenfunction expansions.

To find the eigenvalues and eigenfunctions, seek an exponential solution form $y(x) = e^{rx}$. Putting that form into (31a), we find that $r = 1 \pm \sqrt{1-\lambda}$, so the general solution of (31a) is

$$y(x) = e^x\left(Ae^{\sqrt{1-\lambda}\,x} + Be^{-\sqrt{1-\lambda}\,x}\right), \tag{36}$$

unless $\lambda = 1$, in which case the two solutions in (36) coalesce into one. Thus, we need to distinguish the two cases $\lambda \neq 1$ and $\lambda = 1$, and write the general solution as

$$y(x) = \begin{cases} e^x[C\sinh(\sqrt{1-\lambda}\,x) + D\cosh(\sqrt{1-\lambda}\,x)], & \lambda \neq 1 \\ e^x(E + Fx), & \lambda = 1 \end{cases} \tag{37a,b}$$

where the sinh, cosh combination will be a bit more convenient than the positive and negative exponentials in (36) because the $y(0) = 0$ boundary condition will give $D = 0$ and will thereby knock out one of the two terms.

Applying the boundary conditions to (37b) gives $E = F = 0$, so $\lambda = 1$ is not an eigenvalue of (31). Applying them to (37a) gives $D = 0$ and the characteristic equation

$$\sinh(\sqrt{1-\lambda}\,\pi) = 0, \tag{38}$$

with C remaining arbitrary. For $\lambda < 1$, (38) has no roots. For $\lambda > 1$, write*

$$\sinh(\sqrt{1-\lambda}\,\pi) = \sinh(i\sqrt{\lambda - 1}\,\pi) = i\sin(\sqrt{\lambda-1}\,\pi) = 0, \tag{39}$$

so $\sqrt{\lambda - 1}\,\pi = n\pi$ for $n = 1, 2, \ldots$. Thus, the eigenvalues are

$$\lambda_n = 1 + n^2. \qquad (n = 1, 2, \ldots) \tag{40}$$

Further,

$$\begin{aligned} y(x) &= Ce^x \sinh(\sqrt{1-\lambda}\,x) = Ce^x \sinh(i\sqrt{\lambda - 1}x) \\ &= iCe^x \sin(\sqrt{\lambda - 1}\,x) = iCe^x \sin nx, \end{aligned} \tag{41}$$

so the eigenfunctions are

$$\phi_n(x) = e^x \sin nx. \tag{42}$$

Finally, the eigenfunction expansion of a given function $f(x)$ on $0 < x < \pi$ is

$$f(x) = \sum_{n=1}^{\infty} a_n \phi_n(x), \qquad (0 < x < \pi) \tag{43a}$$

$$\begin{aligned} a_n = \frac{\langle f, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} &= \frac{\int_0^\pi f(x)(e^x \sin nx)e^{-2x}\,dx}{\int_0^\pi (e^x \sin nx)^2 e^{-2x}\,dx} \\ &= \frac{\int_0^\pi f(x)e^{-x}\sin nx\,dx}{\int_0^\pi \sin^2 nx\,dx} = \frac{2}{\pi}\int_0^\pi f(x)e^{-x}\sin nx\,dx. \end{aligned} \tag{43b}$$

COMMENT. We did not use Theorem 17.7.2 in this example because it addresses the distinction $\lambda > 0$, $\lambda < 0$, whereas here we were concerned with the cases $\lambda > 1$, $\lambda < 1$. ∎

In Examples 1 – 4 it turned out that the separately-considered λ's [$\lambda = 0$ in (4), $\lambda = 0$ in (22), and $\lambda = 1$ in (37)] turned out not to be eigenvalues. Do not discard such values out of hand because they might, in other examples, turn out to be eigenvalues. For instance, you will find that if the boundary conditions in Example

*Here we use the definitions of sinh ( ) and sin ( ):

$$\sinh it = \frac{e^{it} - e^{-it}}{2} = i\,\frac{e^{it} - e^{-it}}{2i} = i \sin it.$$

1 are changed to $y'(0) = 0$ and $y'(L) = 0$, then $\lambda = 0$ is indeed an eigenvalue, with the eigenfunction $\phi(x) = 1$.

17.7.2. Lagrange identity and proofs. (Optional) We will derive a Lagrange identity, and use it to prove parts of Theorem 17.7.1. Proof of Theorem 17.7.2 is left for the exercises. We assume elementary knowledge of the Cartesian representation of complex numbers $z = x + iy$ and the complex conjugate $\overline{z} = x - iy$, as covered in Section 21.2.

When we introduced the inner product (14), we noted that if we are to admit complex-valued functions then we should modify the inner product as

$$\langle f, g \rangle = \int_a^b f(x)\, \overline{g}(x)\, w(x)\, dx. \tag{44}$$

That is, we continue to ask $p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ to be real, but it is not at all obvious that assumption implies that the eigenvalues and eigenfunctions must be real. For instance, the real matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

has the complex eigenvalues $\lambda = 1 \pm i$ and complex eigenvectors as well.

To begin, observe that the properties

$$\langle f, g \rangle = \overline{\langle g, f \rangle}, \tag{45a}$$
$$\langle \mu f, g \rangle = \mu \langle f, g \rangle, \tag{45b}$$
$$\langle f, \mu g \rangle = \overline{\mu} \langle f, g \rangle, \tag{45c}$$

follow immediately from (44), where μ is any scalar. Since $\langle f, g \rangle \neq \langle g, f \rangle$ in general, according to (45a), it is sometimes useful to specify that f comes first and g comes second in $\langle f, g \rangle$. In what follows we say that $\langle f, g \rangle$ is g **pre-dotted** with f or f **post-dotted** with g, which terminology is not standard.

Let us express (1a) in operator form as

$$\boxed{L[y] = \lambda y,} \tag{46}$$

where L is the differential operator

$$\boxed{L = -\frac{1}{w}\left[\frac{d}{dx}\left(p\frac{d}{dx}\right) + q\right],} \tag{47}$$

and let u and v be any functions having continuous second derivatives on $[a, b]$ and satisfying the homogeneous boundary conditions (1b). Then

$$\begin{aligned} \langle L[u], v \rangle &= \int_a^b -\frac{1}{w}\left[(pu')' + qu\right]\overline{v} w\, dx \\ &= -\int_a^b \left[(pu')' + qu\right]\overline{v}\, dx, \end{aligned} \tag{48}$$

and integrating the $(pu')'\overline{v}$ term by parts twice, so as to undo the derivatives on u, gives

$$\langle L[u], v\rangle = [p(u\overline{v}' - u'\overline{v})]\Big|_a^b - \int_a^b [(p\overline{v}')' + q\overline{v}]u\, dx. \tag{49}$$

Remembering that $p(x)$, $q(x)$, and $w(x)$ are real, we can re-express the last term in (49) as

$$\begin{aligned} \int_a^b [(p\overline{v}')' + q\overline{v}]u\, dx &= \int_a^b \frac{1}{w}[(p\overline{v}')' + q\overline{v}]u\, w\, dx \\ &= \int_a^b \overline{\frac{1}{w}[(pv')' + qv]}\; u\, w\, dx \\ &= -\int_a^b u\, \overline{L[v]}\, w\, dx \\ &= -\langle u, L[v]\rangle. \end{aligned} \tag{50}$$

Further, the boundary term in (49) is zero because u and v satisfy the homogeneous boundary conditions (1b). For instance, suppose $\alpha = \gamma = 1$ and $\beta = \delta = 0$ in (1b), so $u(a) = 0$, $u(b) = 0$, $v(a) = 0$, and $v(b) = 0$. It follows from the latter two that $\overline{v}(a) = 0$ and $\overline{v}(b) = 0$, so

$$[p(u\overline{v}' - u'\overline{v})]\Big|_a^b = (0 - 0) - (0 - 0) = 0; \tag{51}$$

similarly for any α and β (not both zero) and any γ and δ (not both zero), verification of which claim is left for the exercises. Thus, (49) becomes

$$\boxed{\langle L[u], v\rangle = \langle u, L[v]\rangle,} \tag{52}$$

which formula is known as the **Lagrange identity**.

Proof of Theorem 17.7.1, part (a): Let λ be an eigenvalue of (1) and ϕ a corresponding eigenfunction, so

$$L[\phi] = \lambda\phi. \tag{53}$$

Post-dotting (53) with ϕ and pre-dotting (53) with ϕ gives

$$\langle L[\phi], \phi\rangle = \langle \lambda\phi, \phi\rangle = \lambda\langle \phi, \phi\rangle \tag{54a}$$

from (45b), and

$$\langle \phi, L[\phi]\rangle = \langle \phi, \lambda\phi\rangle = \overline{\lambda}\langle \phi, \phi\rangle \tag{54b}$$

from (45c), respectively. Subtracting (54) from (54a) gives, by virtue of the Lagrange identity,

$$0 = (\lambda - \overline{\lambda})\langle \phi, \phi\rangle = (\lambda - \overline{\lambda})\, \|\phi\|^2\, . \tag{55}$$

The factor $\|\phi\|^2$ is nonzero because ϕ is an eigenfunction, so it follows from (55) that $\lambda - \overline{\lambda} = 0$. Thus, $\overline{\lambda} = \lambda$ so λ is real, as was to be proved. ■

Proof of Theorem 17.7.1, part (b): Let ϕ_j and ϕ_k be eigenfunctions corresponding to distinct eigenvalues λ_j and λ_k, respectively. Thus,

$$L[\phi_j] = \lambda_j \phi_j \qquad \text{and} \qquad L[\phi_k] = \lambda_k \phi_k. \tag{56a,b}$$

If we dot ϕ_k into each side of (56a) and dot each side of (56b) into ϕ_j [i.e., we pre-dot (56a) with ϕ_k and post-dot (56b) with ϕ_j], we obtain

$$\begin{aligned} \langle \phi_k, L[\phi_j] \rangle &= \langle \phi_k, \lambda_j \phi_j \rangle \\ &= \overline{\lambda}_j \langle \phi_k, \phi_j \rangle, \end{aligned} \tag{57}$$

and

$$\begin{aligned} \langle L[\phi_k], \phi_j \rangle &= \langle \lambda_k \phi_k, \phi_j \rangle \\ &= \lambda_k \langle \phi_k, \phi_j \rangle, \end{aligned} \tag{58}$$

respectively. The left-hand sides are equal by virtue of the Lagrange identity, and $\overline{\lambda}_j = \lambda_j$ because the λ's are real, so subtraction of (57) from (58) gives

$$(\lambda_j - \lambda_k) \langle \phi_k, \phi_j \rangle = 0. \tag{59}$$

Finally, $\lambda_j - \lambda_k \neq 0$ by assumption, so it follows that $\langle \phi_k, \phi_j \rangle = 0$, as was to be proved. ■

Before closing this section let us explain the significance of the Lagrange identity (52). The operator L is the differential operator

$$L = \frac{1}{w} \left[\frac{d}{dx} \left(p \frac{d}{dx} \right) + q \right] \tag{60a}$$

on the domain $\mathcal{D}$ of functions, u say, that are defined and have continuous second derivatives on $[a, b]$ and that satisfy the homogeneous boundary conditions*

$$\begin{aligned} \alpha u(a) + \beta u'(a) &= 0, \\ \gamma u(b) + \delta u'(b) &= 0. \end{aligned} \tag{60b}$$

*Note that the definition of the domain $\mathcal{D}$ is part of the definition of L, just as the domain of definition is part of the definition of a function. For instance, the function whose values are $\sin x$ on $0 \leq x \leq \pi$ is different from the function whose values are $\sin x$ on $-\pi \leq x \leq 23$. Thus, the operator is the differential operator (60a) plus the domain of definition. In this discussion, although not elsewhere in this text, we distinguish between the *differential operator* [namely, the "action" (60a)] and *operator* [namely, the action (60a) plus the domain of definition $\mathcal{D}$]. By the way, why do we ask functions $u(x)$ in $\mathcal{D}$ to have continuous second derivatives? They need to be twice differentiable so that $L[u]$ exists because L is a second-order differential operator. Further, we wish to ensure that the inner product integral $\langle L[u], v \rangle$ exists, and we can do that by asking the integrand to be continuous on $[a, b]$. In fact, $L[u]$ will be continuous if u has a continuous second derivative.

More generally, the equation

$$\boxed{\langle L[u], v\rangle = \langle u, L^*[v]\rangle} \tag{61}$$

is used to *define* the **Hermitian conjugate** (or **adjoint**) L^* of the operator L, relative to whatever inner product is chosen. Let us illustrate.

EXAMPLE 5. Find the Hermitian conjugate of the operator consisting of the differential operator

$$L = \frac{d^2}{dx^2} + \frac{d}{dx} + 1 \tag{62a}$$

on the domain $\mathcal{D}$ of real-valued functions defined and having continuous second derivatives on $[0, \pi]$ and satisfying the homogeneous initial conditions

$$u(0) = 0, \qquad u'(0) = 0, \tag{62b}$$

subject to the inner product definition

$$\langle u, v\rangle = \int_0^\pi u(x)v(x)\,dx, \tag{62c}$$

say. Begin with the left-hand side of (61),

$$\langle L[u], v\rangle = \int_0^\pi (u'' + u' + u)v\,dx. \tag{63}$$

Integrating the $u''v$ term by parts twice, the $u'v$ term once, leaving the uv term intact, and using (62b), gives

$$\begin{aligned}\langle L[u], v\rangle &= (u'v - uv' + uv)\Big|_0^\pi + \int_0^\pi u(v'' - v' + v)\,dx \\ &= [u'(\pi) + u(\pi)]v(\pi) - [u(\pi)]v'(\pi) + \langle u, L^*[v]\rangle,\end{aligned} \tag{64}$$

where, from the $v'' - v' + v$ in the integral, we can infer that

$$L^* = \frac{d^2}{dx^2} - \frac{d}{dx} + 1. \tag{65a}$$

To obtain the boundary conditions associated with L^* we see, by comparing (64) with (61), that we need the boundary terms in (64) to drop out. Whereas $u(0) = 0$ and $u'(0) = 0$, the bracketed quantities $u'(\pi) + u(\pi)$ and $u(\pi)$ are not prescribed, so we must have both

$$v(\pi) = 0, \qquad v'(\pi) = 0. \tag{65b}$$

Thus, the Hermitian conjugate operator is the differential operator (65a) on the domain $\mathcal{D}^*$ of real-valued functions defined and having continuous second derivatives on $[0, \pi]$ and satisfying the conditions (65b). ▮

If the operator and its Hermitian conjugate (or adjoint) are identical, then we say that it is **Hermitian** (or **self-adjoint**). Thus, the operator in Example 5 is not

Hermitian because the L^* in (65a) differs from the L in (62a) (by the minus sign in front of the d/dx) and also because the boundary conditions (65b) on functions in $\mathcal{D}^*$ are different from the boundary conditions (62b) on functions in $\mathcal{D}$. Either of these differences would be sufficient to conclude that the operator is not Hermitian.

EXAMPLE 6. *Matrix Case.* Find the Hermitian conjugate of a real $n \times n$ matrix operator $\mathbf{A}$, defined on the vector space $\mathbb{R}^n$, with the dot product $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are n-dimensional column vectors. $\mathbf{A}^*$ is defined by requiring that

$$(\mathbf{Ax}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y}) \tag{66}$$

for all $\mathbf{x}$'s in the domain $\mathcal{D}$ of $\mathbf{A}$ and for all $\mathbf{y}$'s in the domain $\mathcal{D}^*$ of $\mathbf{A}^*$. Since $\mathbf{A}$ is $n \times n$ and $\mathbf{x}$ is $n \times 1$, the $\mathbf{Ax}$ in (66) is $n \times 1$. For the dot product on the left to be defined we need $\mathbf{y}$ to be $n \times 1$ as well. On the right, $\mathbf{x}$ is $n \times 1$, so we need $\mathbf{A}^*\mathbf{y}$ to be $n \times 1$. Since $\mathbf{y}$ is $n \times 1$, $\mathbf{A}^*$ needs to be $n \times n$. Thus, like $\mathbf{A}$, $\mathbf{A}^*$ is an $n \times n$ matrix operator defined on $\mathbb{R}^n$. To determine $\mathbf{A}^*$, begin with the left-hand side of (66),

$$(\mathbf{Ax}) \cdot \mathbf{y} = (\mathbf{Ax})^{\mathrm{T}}\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^{\mathrm{T}}\mathbf{y}), \tag{67}$$

so the Hermitian conjugate $\mathbf{A}^*$ of $\mathbf{A}$ is

$$\mathbf{A}^* = \mathbf{A}^{\mathrm{T}}. \tag{68}$$

Thus, $\mathbf{A}$ is Hermitian if and only if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, that is, if $\mathbf{A}$ is symmetric. Just as the eigenvalue problem for the Hermitian Sturm-Liouville problem is of great importance, so is the eigenvalue problem for real symmetric (hence Hermitian) matrices and, indeed, that case is singled out for study in Section 11.3. ▮

Closure. The Sturm-Liouville eigenvalue problem is the differential equation analog of the matrix eigenvalue problem $\mathbf{Ax} = \lambda\mathbf{x}$, where $\mathbf{A}$ is real and symmetric (hence Hermitian). In both cases the eigenvalues are real and the eigenvectors provide an orthogonal basis for the relevant vector space. The expansion formula (15) is the analog of formula (23) in Section 9.9.

Although the Sturm-Liouville and matrix eigenvalue problems are closely related (if $\mathbf{A}$ is Hermitian), the Sturm-Liouville case is much more subtle because the vector space is infinite-dimensional and expansions are, in general, infinite series. For instance, in a five-dimensional space five orthogonal vectors necessarily constitute a basis, but in an infinite-dimensional space an infinite number of orthogonal vectors need not constitute a basis. Thus, part (d) of Theorem 17.7.1 is a deep result, and we put it forward without proof.

If we were to generalize the real symmetric matrix and Sturm-Liouville eigenvalue problems further, we would study Hermitian operators, with matrix, ordinary differential, partial differential, and integral operators merely occurring as special cases. Such study is beyond our present scope and falls within the domain of mathematics known as functional analysis. Rather than being exceptional, the operators encountered in applications are often Hermitian.

Computer software. To find the roots of (24) using *Maple* we can use the fsolve command, but it is best to include the option of specifying the interval in which to search. From Fig. 3, it is evident that the first $\sqrt{\lambda}$ root falls in the interval $[\pi/2, 3]$. Thus, there is a λ root in $[2.46, 9]$, so enter

$$\text{fsolve}(\tan(\text{sqrt}(x)) + 2 * \text{sqrt}(x) = 0,\ x,\ 2.46..9);$$

and return. The result is the first root, $\lambda = 3.3731$. For the second root enter

$$\text{fsolve}(\tan(\text{sqrt}(x)) + 2 * \text{sqrt}(x) = 0,\ x,\ 22..36);$$

and so on.

EXERCISES 17.7

1. Identify $p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$, solve for the eigenvalues and eigenfunctions, and work out the eigenfunction expansion of the given function f. If the characteristic equation is too difficult to solve analytically, state that and proceed with the rest of the problem as though the λ_n's were known.

(a) $y'' + \lambda y = 0,$
$y(0) = 0, \quad y'(L) = 0, \quad f(x) = 100$

(b) $y'' + \lambda y = 0,$
$y'(0) = 0, \quad y(L) = 0, \quad f(x) = 1$

(c) $y'' + \lambda y = 0,$
$y'(0) = 0, \quad y'(L) = 0, \quad f(x) = \begin{cases} 1, & 0 \le x < L/2 \\ 0, & L/2 \le x \le L \end{cases}$

(d) $y'' + \lambda y = 0,$
$y'(0) = 0, \quad y(L) + y'(L) = 0, \quad f(x) = 50$

(e) $y'' + \lambda y = 0,$
$y(0) + y'(0) = 0, \quad y(\pi) = 0, \quad f(x) = 10$

(f) $y'' + \lambda y = 0,$
$y'(-1) = 0,\ y'(1) = 0,\ f(x) = \begin{cases} 0, & -1 \le x \le 0 \\ 50, & 0 < x \le 1 \end{cases}$

(g) $y'' + \lambda y = 0,$
$y(0) - 2y'(0) = 0, \quad y'(2) = 0, \quad f(x) = 100$

2. We pointed out that the Sturm-Liouville problem in Example 1 generated the half-range sine expansion studied in Section 17.4. Modify the boundary conditions in that example so as to generate, instead, the

(a) half-range cosine expansion
(b) quarter-range sine expansion
(c) quarter-range cosine expansion

3. (*Obtaining Sturm–Liouville form*) We observed, in Example 4, that the equation

$$A(x)y'' + B(x)y' + C(x)y + \lambda D(x)y = 0 \tag{3.1}$$

is in the standard Sturm-Liouville form (1a) only if $B(x) = A'(x)$. Show that if $A(x) \neq 0$ on $[a, b]$ and $(B - A')/A$ is continuous on $[a, b]$, then we can recast (3.1) in the form (1a) by multiplying (3.1) by

$$\sigma(x) = e^{\int[(B-A')/A]\,dx}. \tag{3.2}$$

4. Use the results of Exercise 3 to recast each of the following differential equations in the Sturm–Liouville form (1a). Identify $p(x)$, $q(x)$, and $w(x)$.

(a) $xy'' + 5y' + \lambda xy = 0$
(b) $y'' + 2y' + xy + \lambda x^2 y = 0$
(c) $y'' + y' + \lambda y = 0$
(d) $y'' - y' + \lambda xy = 0$
(e) $x^2y'' + xy' + \lambda x^2 y = 0$
(f) $y'' + (\cot x)y' + \lambda y = 0$

5. Use computer software to find $\lambda_1, \ldots, \lambda_8$ from (24), to seven significant figures.

6. Consider the eigenvalue problem

$$y'' + \lambda y = 0, \qquad (0 < x < 1)$$

$$2y(0) - y(1) + 4y'(1) = 0, \qquad y(0) + 2y'(1) = 0.$$

Explain why the latter problem is not of Sturm-Liouville type. Using computer software, determine any two eigenvalues.

HINT: You should obtain the characteristic equation

$$\sin\sqrt{\lambda} = 2\sqrt{\lambda}.$$

Although the latter equation has no roots on a real $\sqrt{\lambda}$ axis, we need to search in the complex plane. With $z = x + iy$ write $\sin z = 2z$, use the identity $\sin z = \sin(x + iy) = \sin x \cosh y + i \cos x \sinh y$ and obtain the equations $\sin x \cosh y = 2x$, $\cos x \sinh y = 2y$ on x and y. Then, use computer software to find any two solution pairs for x and y, and hence for λ.

7. Show that for

$$\begin{aligned} y'' + \lambda y &= 0, & (0 < x < 1) \\ y(0) - y(1) &= 0, & y'(0) + y'(1) = 0 \end{aligned}$$

every λ (real or complex) is an eigenvalue! Is the latter a Sturm–Liouville system? Explain.

8. Consider the eigenvalue problem

$$x^2 y'' + xy' + \lambda y = 0, \qquad (1 < x < a)$$

$$y(1) = 0, \qquad y(a) = 0.$$

(a) Show that the eigenvalues and eigenfunctions are

$$\lambda_n = \frac{n^2\pi^2}{(\ln a)^2}, \qquad \phi_n(x) = \sin\left(n\pi \frac{\ln x}{\ln a}\right)$$

for $n = 1, 2, \ldots$.

(b) Show that the eigenfunction expansion of a given function f is of the form

$$f(x) = \sum_{n=1}^{\infty} c_n \sin\left(n\pi \frac{\ln x}{\ln a}\right),$$

where

$$c_n = \frac{\displaystyle\int_1^a f(x)\sin\left(n\pi\frac{\ln x}{\ln a}\right)\frac{dx}{x}}{\displaystyle\int_1^a \sin^2\left(n\pi\frac{\ln x}{\ln a}\right)\frac{dx}{x}}.$$

HINT: You will need to get the differential equation into Sturm–Liouville form, as discussed in Exercise 3, before you can identify the weight function $w(x)$ for the inner product.

9. Expand the function

$$f(x) = \begin{cases} x^4, & 0 \le x < 2 \\ 0, & 2 \le x \le \pi \end{cases}$$

in terms of the eigenfunctions of the given eigenvalue problem. Use computer software, such as the *Maple* int command, to evaluate the expansion coefficient a_n as a function of n.

(a) $y'' + \lambda y = 0, \quad y'(0) = 0,\ y'(\pi) = 0$
(b) $y'' + \lambda y = 0, \quad y'(0) = 0,\ y(\pi) = 0$
(c) $y'' + \lambda y = 0, \quad y(0) = 0,\ y'(\pi) = 0$

10. Prove that the Sturm–Liouville eigenvalues are simple, as stated in part (b) of Theorem 17.7.1. HINT: Suppose that ϕ_1 and ϕ_2 are two eigenfunctions corresponding to an eigenvalue λ of (1), and suppose $\alpha \neq 0$ in (1b). Then the Wronskian $W(x)$ of ϕ_1 and ϕ_2, evaluated at $x = a$, is

$$\begin{aligned} W(a) &= \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} \\ &= \begin{vmatrix} -\dfrac{\beta}{\alpha}\phi_1'(a) & -\dfrac{\beta}{\alpha}\phi_2'(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} = 0. \end{aligned} \tag{10.1}$$

On the other hand, if α does equal zero in (1b) then (1b) becomes $y'(0) = 0$, so

$$W(a) = \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} = \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ 0 & 0 \end{vmatrix} = 0 \tag{10.2}$$

once again. Show that it follows from these results and Liouville's formula that $W(x) = 0$ on $[a, b]$, and cite an appropriate theorem which then implies that ϕ_1 and ϕ_2 must be linearly dependent on $[a, b]$.

11. (*Real eigenfunctions*) Show that if $\phi(x)$ is an eigenfunction of a Sturm–Liouville problem, then $\phi(x)$ is either a real-valued function or else it is a complex constant times a real-valued function. HINT: Show that if $\phi(x)$ is an eigenfunction corresponding to an eigenvalue λ, then so is $\overline{\phi}(x)$. Then use part (b) of Theorem 17.7.1 (namely, that the eigenvalues are simple) to show that $\phi(x) = c\overline{\phi}(x)$, where c is a constant. Expressing the latter equation in the (polar) form $A(x)e^{iB(x)} = Ce^{iD}A(x)e^{-iB(x)}$, show that $B(x)$ is, at most, a constant.

12. Prove Theorem 17.7.2. HINT: You may assume that the eigenfunctions are real (proof of which is outlined in Exercise 10). Multiply each term in $(py')' + qy + \lambda wy = 0$ by $\overline{y}$, and integrate over the $[a, b]$ interval. Thus, show that

$$\lambda \|y\|^2 = -(py'\overline{y})\Big|_a^b + \int_a^b p\,|y'|^2\,dx - \int_a^b q\,|y|^2\,dx, \tag{12.1}$$

and examine the signs of the individual terms.

13. (*Column buckling with lateral restraint*) Consider the buckling of the column of length L and stiffness EI shown in the figure. It is fixed into the floor such that $y(0) = y'(0) = 0$, and it is restrained laterally, at the free end, by a spring of stiffness k. Then it turns out that the lateral deflection $y(x)$ is governed by the eigenvalue problem

$$\begin{aligned} &y'''' + \lambda y'' = 0, \qquad (0 < x < L) \\ &y(0) = y'(0) = y''(L) = 0, \qquad\qquad (13.1) \\ &y'''(L) = -\lambda y'(L) + \kappa y(L), \end{aligned}$$

where $\lambda \equiv P/EI$ and $\kappa = k/EI$.

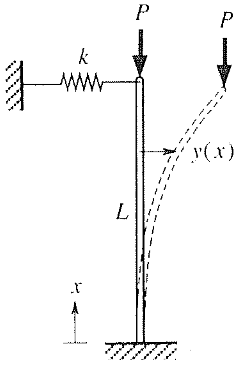
(a) Show that the characteristic equation is

$$(\Lambda^2 - \kappa L)\Lambda \cos \Lambda L + \kappa \sin \Lambda L = 0, \qquad (13.2)$$

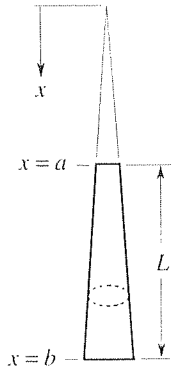
where $\Lambda = \sqrt{\lambda}$, and that the corresponding eigenfunctions (buckling modes) are (to within an arbitrary scale factor)

$$y = \sin \Lambda x - \tan \Lambda L \cos \Lambda x - \Lambda x + \tan \Lambda L. \qquad (13.3)$$

(b) Solve (13.2) for the critical buckling load, P_{cr}, for the case where $\kappa = 0$.
(c) Is (13.1) a Sturm–Liouville problem? Explain.



14. (*Buckling of linearly tapered column*) Consider a column of circular cross section, the radius of which varies linearly with x. It extends over $a < x < b$, as shown in the figure, is pinned at both ends, and is loaded axially by a force P. Then the cross-sectional moment of inertia I is not a constant; it is given by $I(x) = I_0(x/b)^4$, where the constant I_0 is the value of $I(x)$ at $x = b$, so the eigenvalue problem governing buckling is found to be



$$\begin{aligned} &x^4 y'' + \lambda y = 0, \qquad (a < x < b) \\ &y(a) = 0, \quad y(b) = 0, \end{aligned} \qquad (14.1)$$

where $\lambda \equiv b^4 P/EI_0$.

(a) Verify that (for the case $\lambda \neq 0$)the general solution of the differential equation can be expressed as

$$y(x) = x\left[A \sin\left(\frac{\sqrt{\lambda}}{x}\right) + B \cos\left(\frac{\sqrt{\lambda}}{x}\right)\right]. \qquad (14.2)$$

(b) Applying the boundary conditions, show that the eigenvalues and eigenfunctions are

$$\lambda_n = \left(\frac{n\pi ab}{L}\right)^2, \qquad (14.3a)$$

$$\phi_n(x) = x \sin\left[n\pi \frac{b}{L}\left(1 - \frac{a}{x}\right)\right] \qquad (14.3b)$$

for $n = 1, 2, \ldots$, and that the buckling load is $P_{cr} = \pi^2 EI_0 a^2/b^2 L^2$, where $L = b - a$.

15. (*Buckling of nonlinearly tapered column*) Although not wishing to give undue prominence to the subject of the buckling of columns, we include this final exercise on buckling, which we believe is interesting and challenging. If, in the problem of Exercise 14, the column radius is proportional to $\sqrt{x}$ rather than to x, then $I(x) = I_0(x/b)^2$, and we have

$$\begin{aligned} &x^2 y'' + \lambda y = 0, \qquad (a < x < b) \\ &y(a) = 0, \quad y(b) = 0, \end{aligned} \qquad (15.1)$$

where $\lambda \equiv b^2 P/EI_0$. Show that the buckling load is

$$P_{cr} = \frac{EI_0}{b^2}\left\{\frac{1}{4} + \frac{\pi^2}{[\ln(b/a)]^2}\right\}. \qquad (15.2)$$

16. We show in (51) that the boundary term in (49) is zero for the special simple case where $\alpha = \gamma = 1$ and $\beta = \delta = 0$. Prove that the boundary term is zero for *any* α and β (not both zero) and for *any* γ and δ (not both zero).

17. Find the Hermitian conjugate (i.e., the adjoint) of the given operator and state whether the given operator is Hermitian (self-adjoint). If it is not, state why it is not. In each case, the interval is $0 \le x \le 1$, and the inner product is

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx.$$

(a) $L = \dfrac{d}{dx}, \quad u(0) = 0$

(b) $L = \dfrac{d}{dx}, \quad u(1) = 0$

(c) $L = \dfrac{d^2}{dx^2}, \quad u(0) = u'(0) = 0$

(d) $L = \dfrac{d^2}{dx^2}, \quad u'(0) = u'(1) = 0$

(e) $L = \dfrac{d^2}{dx^2} + 3, \quad u'(0) = u'(1) = 0$

(f) $L = \dfrac{d^2}{dx^2} + \dfrac{d}{dx}, \quad u(0) = u(1) = 0$

(g) $L = \dfrac{d^3}{dx^3} - \dfrac{d^2}{dx^2} + 2\dfrac{d}{dx}, \quad u(0) = u'(0) = u'(1) = 0$

(h) $L = \dfrac{d^2}{dx^2} - 1, \quad u(0) + u'(0) = u'(1) = 0$

(i) $L = \dfrac{d^2}{dx^2}, \quad u'(0) = u(1) + 5u'(1) = 0$

(j) $L = \dfrac{d^2}{dx^2}, \quad u(0) - u(1) = u'(0) - u'(1) = 0$

17.8 Periodic and Singular Sturm–Liouville Problems

The Sturm–Liouville problem studied in Section 17.7 consists of the linear homogeneous second-order differential equation

$$[p(x)y']' + q(x)y + \lambda w(x)y = 0, \qquad (a < x < b) \qquad (1a)$$

with homogeneous boundary conditions of the form

$$\begin{aligned} \alpha y(a) + \beta y'(a) &= 0,\\ \gamma y(b) + \delta y'(b) &= 0, \end{aligned} \qquad (1b)$$

where a, b are finite, where p, p', q, w are continuous on $[a, b]$, and where $p(x) > 0$ and $w(x) > 0$ on $[a, b]$. The latter is generally known as a **regular** Sturm–Liouville problem, and many powerful results followed, as given in Theorems 17.7.1 and 17.7.2. We say that the boundary conditions (1b) are **separated** since one condition applies at $x = a$ and the other at $x = b$.

If any of the conditions cited above are not met, then the results obtained in Theorems 17.7.1 and 17.7.2 may not hold. Among the nonregular versions of the Sturm–Liouville problem, two are especially prominent and are the subject of this section: the Sturm–Liouville problem with **periodic boundary conditions**, and the **singular** Sturm–Liouville problem. In these cases the conditions cited above are met, except as noted below.

Periodic boundary conditions. In this case we have, in place of the separated boundary conditions (1b), the nonseparated conditions

$$\boxed{\begin{aligned} y(a) &= y(b), \\ y'(a) &= y'(b), \end{aligned}} \tag{2}$$

which are known as *periodic boundary conditions*, for reasons that will become clear when we work an example.

Singular case. In this case $p(x)$ [and possibly $w(x)$] vanishes at one or both endpoints, so that $p(x) > 0$ and $w(x) > 0$ holds on the open interval (a, b) rather than on the closed interval $[a, b]$. Further, the boundary conditions are modified as follows.

$\boldsymbol{p(a) = 0}$ [and $p(b) \neq 0$]: Then the boundary conditions are

$$\boxed{\begin{aligned} &y \text{ bounded at } a, \\ &\gamma y(b) + \delta y'(b) = 0. \end{aligned}} \tag{3}$$

$\boldsymbol{p(b) = 0}$ [and $p(a) \neq 0$]: Then the boundary conditions are

$$\boxed{\begin{aligned} &\alpha y(a) + \beta y'(a) = 0, \\ &y \text{ bounded at } b. \end{aligned}} \tag{4}$$

$\boldsymbol{p(a) = p(b) = 0}$: Then the boundary conditions are

$$\boxed{\begin{aligned} &y \text{ bounded at } a, \\ &y \text{ bounded at } b. \end{aligned}} \tag{5}$$

By y being bounded at a, for example, we mean that $\lim_{x \to a} y(x)$ exists (and is therefore finite).

For these cases we have the following results.

THEOREM 17.8.1 *Periodic and Singular Cases*
Let λ_n and $\phi_n(x)$ denote any eigenvalue and corresponding eigenfunction of a Sturm–Liouville problem with periodic boundary conditions, given by (2), or a singular Sturm–Liouville problem (as defined above).

(a) The eigenvalues are real.

(b) If $q(x) \leq 0$ on $[a,b]$ and $[p(x)\phi_n(x)\phi_n'(x)]|_a^b \leq 0$ for the eigenfunction $\phi_n(x)$, then not only is λ_n real, it is also nonnegative: $\lambda_n \geq 0$.

(c) Eigenfunctions corresponding to distinct eigenvalues are orthogonal. That is, if $\lambda_j \neq \lambda_k$, then $\langle \phi_j, \phi_k \rangle = 0$.

As for the regular case, positive statements can be made about the completeness of the sets of orthogonal eigenfunctions generated by these problems, in the sense of their being bases for the eigenfunction expansion representation of sufficiently well behaved functions on the interval $a < x < b$.

EXAMPLE 1. *Periodic Boundary Conditions.* Consider the Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad (-L < x < L) \tag{6a}$$

$$y(-L) = y(L), \quad y'(-L) = y'(L). \tag{6b}$$

We begin with the general solution

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx, & \lambda = 0. \end{cases} \tag{7a,b}$$

For $\lambda = 0$ the boundary conditions (6b) give $C - DL = C + DL$ and $D = D$, so $D = 0$ and C is arbitrary. Thus, $y(x) = C$, so $\lambda = 0$ is an eigenvalue and has the eigenfunction $\phi(x) = 1$. For $\lambda \neq 0$ the boundary conditions (6b) give

$$(\sin\sqrt{\lambda}\,L)B = 0 \tag{8a}$$

$$(\sin\sqrt{\lambda}\,L)A = 0. \tag{8b}$$

Thus, either $A = B = 0$, which result we reject because it gives only the trivial solution $y(x) = 0$, or else

$$\sin\sqrt{\lambda}\,L = 0 \tag{9}$$

and A, B are arbitrary. Since $q(x) = 0$, and

$$\begin{aligned}[p(x)\phi_n(x)\phi_n'(x)]\Big|_{-L}^{L} &= \phi_n(L)\phi_n'(L) - \phi_n(-L)\phi_n'(-L) \\ &= \phi_n(L)\phi_n'(L) - \phi_n(L)\phi_n'(L) = 0,\end{aligned}$$

it follows from part (b) of Theorem 17.8.1 that $\lambda_n \geq 0$. Thus, $\sqrt{\lambda}$ in (9) is real and (9) has the roots $\sqrt{\lambda}\,L = n\pi$, so the eigenvalues and eigenfunctions are

$$\lambda_0 = 0, \qquad \phi_0(x) = 1 \tag{10a}$$

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \qquad \phi_n(x) = \cos\frac{n\pi x}{L} \text{ and } \sin\frac{n\pi x}{L} \tag{10b}$$

for $n = 1, 2, \ldots$.

Remember that there is nothing inappropriate about an eigenvalue being zero. It is the eigen*function* that is to be nontrivial, and $\lambda_0 = 0$ does give the nontrivial solution $\phi_0(x) = 1$.

Observe that the eigenvalues $\lambda_1, \lambda_2, \ldots$ are nonsimple since each one has two linearly independent eigenfunctions. This result could not have occurred in a regular Sturm–Liouville problem, which must have simple eigenvalues [part (b) of Theorem 17.7.1], and is due to the periodic boundary conditions.

Noting that the weight function is $w(x) = 1$, the eigenfunction expansion of a given function f on $(-L, L)$ takes the form

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right), \tag{11a}$$

$$a_0 = \frac{\langle f, 1 \rangle}{\langle 1, 1 \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x)\, 1\, dx}{\displaystyle\int_{-L}^{L} 1^2\, dx} = \frac{1}{2L} \int_{-L}^{L} f(x)\, dx, \tag{11b}$$

$$\begin{aligned} a_n &= \frac{\langle f, \cos \frac{n\pi x}{L} \rangle}{\langle \cos \frac{n\pi x}{L}, \cos \frac{n\pi x}{L} \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx}{\displaystyle\int_{-L}^{L} \cos^2 \frac{n\pi x}{L}\, dx} \\ &= \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx, \end{aligned} \tag{11c}$$

$$\begin{aligned} b_n &= \frac{\langle f, \sin \frac{n\pi x}{L} \rangle}{\langle \sin \frac{n\pi x}{L}, \sin \frac{n\pi x}{L} \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx}{\displaystyle\int_{-L}^{L} \sin^2 \frac{n\pi x}{L}\, dx} \\ &= \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx. \end{aligned} \tag{11d}$$

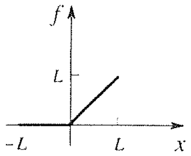


Figure 1. Graph of f.

The result (11) is seen to be the same as the classical Fourier series of a $2L$-periodic function, which we studied in Section 17.3. Thus the expansion (11a) of the function f shown in Fig. 1, for example, will be the same as the classical Fourier series of the $2L$-periodic function shown in Fig. 2, which result explains why we call (2) "periodic boundary conditions."

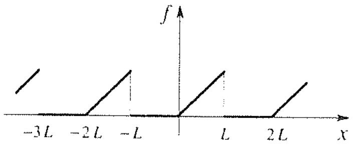


Figure 2. $2L$-periodic f.

COMMENT. As usual, the function f being expanded (e.g., f in Fig. 1) does not itself need to satisfy the boundary conditions imposed on the eigenfunctions [(6b) in this case]. ▮

EXAMPLE 2. *A Bessel Equation.* Consider the singular Sturm–Liouville problem

$$(xy')' + \lambda xy = 0, \qquad (0 < x < L) \tag{12a}$$

$$y(0) \text{ bounded}, \quad y(L) = 0. \tag{12b}$$

Since (12a) is already in the standard Sturm–Liouville form $(py')' + qy + \lambda wy = 0$, we can see that $p(x) = x$, $q(x) = 0$, and $w(x) = x$. And since $p(x)$ and $w(x)$ vanish at the left endpoint $x = 0$, the problem (12) is singular; hence the boundary condition adopted at $x = 0$ is simply a boundedness condition.

To solve (12a), we notice that (12a) would be a Bessel equation of order zero if the λ were not present. Let us try to convert (12a) to a Bessel equation by a simple scaling, $x = \alpha t$, where α is to be determined. Under that change of variables (12a) becomes

$$(tT')' + \alpha^2 \lambda t T = 0, \tag{13}$$

where $T(t) \equiv y(x(t)) = y(\alpha t)$, and where the primes denote d/dt. Thus, we can remove the λ by choosing $\alpha = 1/\sqrt{\lambda}$ (tentatively assuming that $\lambda \neq 0$), so $t = \sqrt{\lambda}\, x$. Then (13) becomes $(tT')' + tT = 0$, with the general solution $T(t) = AJ_0(t) + BY_0(t)$. Or, reverting to x and y,

$$y(x) = AJ_0(\sqrt{\lambda}\, x) + BY_0(\sqrt{\lambda}\, x). \tag{14}$$

However, for $\lambda = 0$ the latter fails to provide the general solution of (12a) because $Y_0(0) = -\infty$ is undefined. But if $\lambda = 0$ then (12a) becomes $(xy')' = 0$ which can be integrated to give $y(x) = C + D \ln x$. Thus, let us write the general solution of (12a) as

$$y(x) = \begin{cases} AJ_0(\sqrt{\lambda}\, x) + BY_0(\sqrt{\lambda}\, x), & \lambda \neq 0 \\ C + D \ln x, & \lambda = 0. \end{cases} \tag{15a,b}$$

For $\lambda = 0$, the boundedness condition requires that $D = 0$, and then $y(L) = 0$ gives $C = 0$, so $y(x) = 0$. Therefore, $\lambda = 0$ is not an eigenvalue. For $\lambda \neq 0$, the boundedness condition requires that $B = 0$ (because $Y_0 \to -\infty$ as its argument tends to zero), so

$$y(x) = AJ_0(\sqrt{\lambda}\, x). \tag{16}$$

Then, the other boundary condition gives

$$y(L) = 0 = AJ_0(\sqrt{\lambda}\, L). \tag{17}$$

If $A = 0$, then (16) becomes the trivial solution, so let us satisfy (17) by asking that

$$J_0(\sqrt{\lambda}\, L) = 0 \tag{18}$$

instead. Now, $q(x) = 0$, and

$$[p(x)\phi_n(x)\phi_n'(x)]\Big|_0^L = [x\phi_n(x)\phi_n'(x)]\Big|_0^L \tag{19}$$

is zero because $\phi_n(L) = 0$ and the x factor is zero at $x = 0$,[*] so Theorem 17.8.1 tells us that not only is λ real, it is also nonnegative. Thus, the argument of J_0 in (18) is real, and it suffices to look for roots of (18) on the real axis. If we denote the zeros of $J_0(x)$ as $x = z_1, z_2, \ldots$ (Fig. 3), then (18) gives $\sqrt{\lambda}\, L = z_n$, so the eigenvalues and eigenfunctions of (12) are

$$\lambda_n = \left(\frac{z_n}{L}\right)^2 \quad \text{and} \quad \phi_n(x) = J_0\left(z_n \frac{x}{L}\right) \tag{20}$$

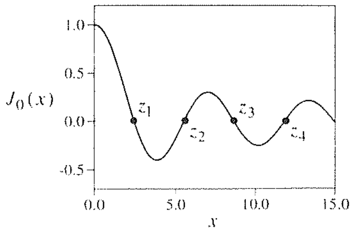


Figure 3. The zeros z_n of J_0.

[*] We can conclude that $x\phi_n\phi_n'$ vanishes at $x = 0$, provided that ϕ_n and ϕ_n' are finite there. We know that ϕ_n is finite there because that is our boundary condition, at $x = 0$, in (12b). Thus, we need to also ask that ϕ_n' be finite there. But by the time we reach (19) we already have the form (16) for the eigenfunctions, even if we don't yet know the λ's, and the derivative of the right-hand side of (16) is bounded at $x = 0$; indeed, it is even zero.

for $n = 1, 2, \dots$. The zeros z_n are tabulated, and the first several are as follows:

$$z_1 = 2.405, \quad z_2 = 5.520, \quad z_3 = 8.654, \quad z_4 = 11.792, \quad z_5 = 14.931, \quad \dots. \tag{21}$$

Further, the eigenfunction expansion of a given function f, on $0 < x < L$, is given by

$$f(x) = \sum_{n=1}^{\infty} a_n J_0 \left(z_n \frac{x}{L} \right), \tag{22}$$

where, recalling that the weight function in the inner product is $w(x) = x$,

$$a_n = \frac{\langle f(x), J_0 \left(z_n \frac{x}{L} \right) \rangle}{\langle J_0 \left(z_n \frac{x}{L} \right), J_0 \left(z_n \frac{x}{L} \right) \rangle} = \frac{\int_0^L f(x) J_0 \left(z_n \frac{x}{L} \right) x \, dx}{\int_0^L \left[J_0 \left(z_n \frac{x}{L} \right) \right]^2 x \, dx}. \tag{23}$$

The integral in the denominator of (23) is evaluated in Exercise 7 of Section 4.6, so the final expression for a_n is

$$a_n = \frac{2}{L^2 [J_1(z_n)]^2} \int_0^L f(x) J_0 \left(z_n \frac{x}{L} \right) x \, dx, \tag{24}$$

where J_0 and J_1 are Bessel functions of the first kind, of orders 0 and 1, respectively. ▮

EXAMPLE 3. *A Legendre Equation.* The Sturm–Liouville problem

$$(1 - x^2) y'' - 2xy' + \lambda y = 0 \qquad (-1 < x < 1) \tag{25a}$$

$$y(-1) \text{ bounded}, \quad y(1) \text{ bounded} \tag{25b}$$

is also singular because $p(x) = 1 - x^2$ vanishes at both endpoints. Hence, we apply the boundedness boundary conditions. In fact, (25a) is the Legendre equation, which is the subject of Section 4.4. There, we found that solutions of (25a) that are bounded on $-1 \le x \le 1$ are possible only if $\lambda = n(n+1)$, for $n = 0, 1, 2, \dots$, and those nontrivial solutions are the Legendre polynomials $P_n(x)$. Thus, the eigenvalues and eigenfunctions of (25) are

$$\lambda_n = n(n+1), \qquad \phi_n(x) = P_n(x), \tag{26}$$

for $n = 0, 1, 2, \dots$.

The eigenfunction expansion of a given function f, on $-1 \le x \le 1$, is given by

$$f(x) = \sum_{n=0}^{\infty} a_n P_n(x), \tag{27}$$

where, since the weight function is $w(x) = 1$,

$$a_n = \frac{\langle f(x), P_n(x) \rangle}{\langle P_n(x), P_n(x) \rangle} = \frac{\int_{-1}^{1} f(x) P_n(x) \, dx}{\int_{-1}^{1} P_n^2(x) \, dx}. \tag{28}$$

The integral in the denominator of (28) was found [(18) in Section 4.4] to be $2/(2n+1)$, so

$$a_n = \frac{2n+1}{2} \int_{-1}^{1} f(x) P_n(x)\, dx. \tag{29}$$

For instance, let f be the "ramp" $f(x) = xH(x)$, where H is the Heaviside function. Then (29) becomes

$$a_n = \frac{2n+1}{2} \int_{0}^{1} x P_n(x)\, dx, \tag{30}$$

so

$$a_0 = \frac{1}{2} \int_0^1 x\, dx = \frac{1}{4},$$
$$a_1 = \frac{3}{2} \int_0^1 x^2\, dx = \frac{1}{2},$$

and so on. Thus,

$$f(x) = \frac{1}{4} P_0(x) + \frac{1}{2} P_1(x) + \frac{5}{16} P_2(x) - \frac{3}{32} P_4(x) + \frac{13}{256} P_6(x) - \cdots. \tag{31}$$

For comparison, we have plotted both $f(x)$ and the partial sum $s_n(x)$ of the first n nonvanishing terms on the right-hand side of (31), for $n = 2$ and $n = 5$, in Fig. 4. ▮

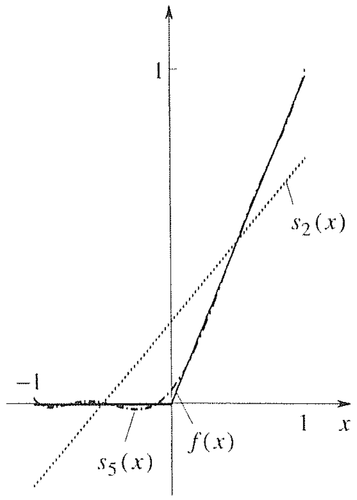


Figure 4. Convergence of (31) to $f(x)$.

Because of their close relationship with Fourier series, we call (22) and (31) **Fourier–Bessel** and **Fourier–Legendre** series, respectively. Such expansions will be needed in Chapters 18–20.

Closure. We have studied the Sturm–Liouville problem with periodic boundary conditions [specifically, the nonseparated conditions (2)], and the singular Sturm–Liouville problem [where $p(x) > 0$ and $w(x) > 0$ hold on $a < x < b$ rather than on $a \leq x \leq b$, and where a boundedness boundary condition is imposed at an endpoint at which $p(x)$ vanishes] because those cases are not covered by the theorems given in Section 17.7, and because they are important cases. Essentially, the upshot is that "all is well": we still obtain real eigenvalues and sets of orthogonal eigenfunctions that can be used to expand functions over the (a, b) interval.

Observe that for the interval $0 < x < 1$, say, each of the Sturm–Liouville problems

$$y'' + \lambda y = 0, \qquad y(0) = 0,\ y(1) = 0$$

and

$$(1 - x^2)y'' - 2xy' + \lambda y = 0, \qquad y'(0) = 0,\ y(1) = \text{bounded}$$

generates an orthogonal basis of eigenfunctions, and that one could write down an unlimited number of other Sturm–Liouville problems that generate orthogonal bases over the same interval. If we wish to expand a given function on that interval, then how do we know which basis to use? As we shall see in Chapters 18–20, that decision will be based on the mathematical context.

EXERCISES 17.8

1. We derived the solution (14) of (12a) by introducing a change of variables $x = t/\sqrt{\lambda}$. Derive it using the method explained in Section 4.6.6, instead.

2. Find the eigenvalues and eigenfunctions, and work out the eigenfunction expansion of the given f. NOTE: As usual, $H(x)$ denotes the Heaviside step function. Use computer software, if you wish, to evaluate any needed integrals.

(a) $y'' + \lambda y = 0, \quad y(0) = y(4),\ y'(0) = y'(4),\ f(x) = H(x-2)$
(b) $y'' + \lambda y = 0, \quad y(-1) = y(5),\ y'(-1) = y'(5),\ f(x) = x + 2$
(c) $x^2y'' + xy' + \lambda y = 0, \quad y(1) = y(2),\ y'(1) = y'(2),\ f(x) = 6$
(d) $(1 - x^2)y'' - 2xy' + \lambda y = 0, \quad y(0) = 0$, $y(1)$ bounded, $f(x) = 4$; evaluate only the first three nonvanishing terms in the expansion of f
(e) $(1 - x^2)y'' - 2xy' + \lambda y = 0, \quad y'(0) = 0$, $y(1)$ bounded, $f(x) = x$; evaluate only the first three nonvanishing terms in the expansion of f
(f) $(1 - x^2)y'' - 2xy' + \lambda y = 0, \quad y(-1)$ bounded, $y'(0) = 0,\ f(x) = 5x^2$
(g) $(4 - x^2)y'' - 2xy' + \lambda y = 0, \quad y(-2)$ bounded, $y(2)$ bounded, $f(x) = 5 - 2x$

3. Expand $f(x) = H(x)$, on $-1 < x < 1$, in terms of the eigenfunctions of the Sturm–Liouville problem

$$(1 - x^2)y'' - 2xy' + \lambda y = 0,$$

where $y(-1)$ and $y(1)$ are bounded. Plot both $f(x)$ and the sum of the first four nonvanishing terms of that expansion.

4. Expand $f(x) = 1 - x$ on $0 < x < 1$, in terms of the eigenfunctions of the Sturm–Liouville problem

$$(1 - x^2)y'' - 2xy' + \lambda y = 0,$$

where $y'(0) = 0$ and $y(1)$ is bounded. Plot both $f(x)$ and the sum of the first three nonvanishing terms of that expansion.

5. Determine the eigenvalues (or at least the characteristic equation for them), eigenfunctions, and weight function of the Sturm–Liouville problem

$$x^2y'' + xy' + (\lambda x^2 - 9)y = 0,$$

where $y(0)$ is bounded and $y(L) = 0$.

6. (*Chebyshev polynomials*) Consider the eigenvalue problem

$$\boxed{(1 - x^2)y'' - xy' + \lambda y = 0, \qquad (-1 < x < 1)} \tag{6.1}$$

where $y(-1)$, $y'(-1)$, $y(1)$, and $y'(1)$ are to be bounded; (6.1) is the **Chebyshev equation**, after the Russian mathematician *Pafnuti Chebyshev* (1821–1894), often transliterated as Tchebichef.

(a) Show that under the change of variables $x = \cos\theta$ the equation (6.1) becomes

$$\Theta'' + \lambda\Theta = 0, \qquad (0 < \theta < \pi) \tag{6.2}$$

where $\Theta(\theta) \equiv y(x(\theta)) = y(\cos\theta)$. Thus, the general solution, in terms of θ, is

$$\Theta(\theta) = \begin{cases} A\cos\sqrt{\lambda}\,\theta + B\sin\sqrt{\lambda}\,\theta, & \lambda \neq 0 \\ C + D\theta, & \lambda = 0. \end{cases} \tag{6.3}$$

(b) Surely, the solutions $\cos\sqrt{\lambda}\,\theta$, $\sin\sqrt{\lambda}\,\theta$, 1, and θ in (6.3) are bounded at $\theta = 0$ $(x = 1)$ and $\theta = \pi$ $(x = -1)$. However, show (by chain differentiation) that $y'(x)$ is bounded at $x = \pm 1$ only if $B = D = 0$ and $\sqrt{\lambda} = n = 1, 2, \ldots$. Thus, the eigenfunctions of (6.1) are, in terms of θ, $\cos n\theta$ $(n = 1, 2, \ldots)$ and 1 or, equivalently, $\cos n\theta$ for $n = 0, 1, 2, \ldots$. In terms of the original x variable, the eigenvalues and eigenfunctions of (6.1) are

$$\boxed{\lambda_n = n^2, \quad T_n(x) = \cos(n\cos^{-1}x), \quad (n = 0, 1, 2, \ldots)} \tag{6.4}$$

the T in honor of Chebyshev.

(c) Though not obvious, it turns out that $T_n(x)$ is an nth-degree polynomial in x. Show that the first several are as follows:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ T_5(x) &= 16x^5 - 20x^3 + 5x. \end{aligned} \tag{6.5}$$

HINT: Use the trigonometric identities

$$\begin{aligned} \cos 2\theta &= 2\cos^2\theta - 1, \\ \cos 3\theta &= 4\cos^3\theta - 3\cos\theta, \\ \cos 4\theta &= 8\cos^4\theta - 8\cos^2\theta + 1, \\ \cos 5\theta &= 16\cos^5\theta - 20\cos^3\theta + 5\cos\theta, \end{aligned}$$

and so on.

(d) Get (6.1) into the standard Sturm–Liouville form by multiplying through by a suitably chosen function $\sigma(x)$ (that is, nonzero on $-1 < x < 1$), and thus show that the weight function is

$$w(x) = \frac{1}{\sqrt{1-x^2}}. \tag{6.6}$$

(e) Theorem 17.8.1 guarantees that, with respect to the weight function (6.6), $\langle T_m(x), T_n(x)\rangle = 0$ for $m \neq n$. Nonetheless, prove that result directly, by evaluating the integral

$$\langle T_m, T_n\rangle = \int_{-1}^{1} T_m(x)T_n(x)\,\frac{1}{\sqrt{1-x^2}}\,dx.$$

Further, show by direct integration that for $m = n$ we have

$$\langle T_n, T_n\rangle = \int_{-1}^{1} T_n^2(x)\,\frac{1}{\sqrt{1-x^2}}\,dx = \begin{cases} \pi, & m = n = 0 \\ \dfrac{\pi}{2}, & m = n \neq 0. \end{cases} \tag{6.7}$$

(f) Thus, the eigenfunction expansion of a given function f, defined on $-1 < x < 1$, is

$$\boxed{f(x) = \sum_{n=0}^{\infty} a_n T_n(x), \qquad (-1 < x < 1)} \tag{6.8}$$

where

$$a_n = \frac{\langle f, T_n\rangle}{\langle T_n, T_n\rangle} = \begin{cases} \dfrac{1}{\pi}\displaystyle\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx, & n = 0 \\ \dfrac{2}{\pi}\displaystyle\int_{-1}^{1} \frac{f(x)T_n(x)}{\sqrt{1-x^2}}\,dx, & n = 1, 2, \dots. \end{cases} \tag{6.9}$$

Use (6.9) to evaluate the a_n's for the case where $f(x) = H(x)$.

(g) Plot $H(x)$ and (by computer) the partial sum of the series obtained in part (f), through $n = 5$.

(h) Derive the values

$$T_n(1) = 1, \quad T_n(-1) = (-1)^n, \tag{6.10}$$

and the recursion formula

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{6.11}$$

(i) Use (6.11), and the T_n's given in (6.5), to derive $T_6(x)$ and $T_7(x)$.

(j) It can be shown that

$$\frac{1 - xt}{1 - 2xt + t^2} = \sum_{n=0}^{\infty} T_n(x)t^n, \qquad (-1 < t < 1) \tag{6.12}$$

so the left-hand member of (6.12) is called a **generating function** of the T_n's. By working out the Taylor series of the left-hand side, verify (6.12) through $n = 2$.

17.9 Fourier Integral

If a function f defined on $-\infty < x < \infty$ is periodic (and sufficiently well-behaved), then it can be represented by a Fourier series. We have begun to see, and will continue to see in Chapters 18–20, that Fourier series representation is of great importance. Sometimes we work with functions, defined on $-\infty < x < \infty$, that are not periodic, such as $f(x) = e^{-x^2}$, the graph of which is given in Fig. 1.

Evidently, we cannot expand such functions in Fourier series if they are not periodic. Yet, we can think of f as periodic but with an infinite period.

Thus, to extend the Fourier series concept to nonperiodic functions we will now consider the limiting case of the classical Fourier series *

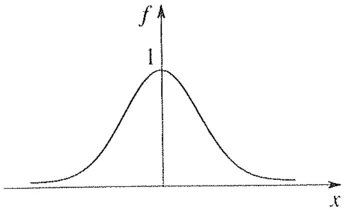


Figure 1. Graph of $f(x) = e^{-x^2}$.

$$f(x) = \sum_{n=0}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \tag{1a}$$

where

$$a_0 = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\,dx, \qquad a_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \cos \frac{n\pi x}{\ell}\,dx,$$
$$b_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \sin \frac{n\pi x}{\ell}\,dx, \tag{1b}$$

as $\ell \to \infty$. We cannot simply *set* $\ell = \infty$ in (1), as that would yield nonsense. Rather one needs to carry out a careful limit process, as ℓ *tends* to ∞.

First, note that the $n\pi/\ell$'s in (1a) are the *frequencies* – spatial or temporal, depending on whether x is a space variable or time. The set of all of the frequencies,

$$0, \frac{\pi}{\ell}, \frac{2\pi}{\ell}, \frac{3\pi}{\ell}, \ldots$$

is called the **frequency spectrum**. To see what happens to the frequency spectrum as ℓ increases, consider the cases where $\ell = \pi$, 2π, and 10π. The corresponding frequency spectra are as follows:

$$\begin{aligned} \ell = \pi : &\qquad n\pi/\ell = 0, 1, 2, 3, 4, \ldots \\ \ell = 2\pi : &\qquad n\pi/\ell = 0, 0.5, 1.0, 1.5, 2.0, \ldots \\ \ell = 10\pi : &\qquad n\pi/\ell = 0, 0.1, 0.2, 0.3, 0.4, \ldots \end{aligned}$$

Observe that as ℓ increases the discrete spectrum becomes more and more dense, and approaches a *continuous spectrum* (from 0 to ∞) as $\ell \to \infty$. Therefore, we can expect that as $\ell \to \infty$ the summation in (1a), on the discrete variable n, will give way to an integration on a continuous variable, say ω. In fact, if f is sufficiently well-behaved (e.g., see Theorem 17.9.1 below) one can show that

$$\boxed{f(x) = \int_0^{\infty} [a(\omega) \cos \omega x + b(\omega) \sin \omega x]\,d\omega,} \tag{2a}$$

*Although we usually split out the a_0 term, here we include it in the sum, merely to increase the resemblance between (1a) and (2a), below.

where[*]

$$\boxed{\begin{aligned} a(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x\,dx, \\ b(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x\,dx. \end{aligned}} \tag{2b}$$

The right-hand side of (2a) is called the **Fourier integral** of f, which we denote as FI f, and (2a) is called the Fourier integral representation of f.

THEOREM 17.9.1 *Fourier Integral Theorem*
Let f be defined on $-\infty < x < \infty$, let f and f' be piecewise continuous on every finite interval $[-\ell, \ell]$ (i.e., for ℓ arbitrarily large), and let $\int_{-\infty}^{\infty} |f(x)|\,dx$ be convergent. Then the Fourier integral of f converges to $f(x)$ at every point x at which f is continuous, and to the mean value $[f(x+) + (f(x-)]/2$ at every point x at which f is discontinuous.

Proof: Rigorous proof of this theorem is well beyond our present scope, and the following is put forward only as a heuristic derivation. If we change the dummy integration variable to ξ in (1b), to distinguish it from the fixed point x in (1a), insert (1b) into (1a), and use the identity $\cos A\cos B + \sin A\sin B = \cos(A - B)$ for greater compactness, then the right-hand side of (1a), namely, FS f, becomes

$$\text{FS}\, f = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(\xi)\,d\xi + \sum_{n=1}^{\infty}\frac{1}{\ell}\int_{-\ell}^{\ell} f(\xi)\cos\frac{n\pi}{\ell}(\xi - x)\,d\xi. \tag{3}$$

Denote the two terms on the right-hand side as I and J, respectively. For I we have

$$|I| = \left|\frac{1}{2\ell}\int_{-\ell}^{\ell} f(\xi)\,d\xi\right| \le \frac{1}{2\ell}\int_{-\infty}^{\infty} |f(\xi)|\,d\xi \to 0 \tag{4}$$

[*] Normally, a singular integral of the form $\int_{-\infty}^{\infty} F(x)\,dx$ is understood to mean

$$\int_{-\infty}^{\infty} F(x)\,dx = \lim_{A\to\infty, B\to\infty}\int_{-A}^{B} F(x)\,dx,$$

where A and B tend to infinity independently. However, the integrals in (2b) are to be understood in the more forgiving sense,

$$\int_{-\infty}^{\infty} F(x)\,dx = \lim_{A\to\infty}\int_{-A}^{A} F(x)\,dx.$$

To see that the latter is "more forgiving" than the former, observe that if $F(x) = x$, for instance, then with the former interpretation the integral is divergent but with the latter interpretation it converges to zero.

as $\ell \to \infty$ because $\int_{-\infty}^{\infty} |f(\xi)|\, d\xi$ is finite, by assumption, and $1/(2\ell) \to 0$ as $\ell \to \infty$. Thus, $I \to 0$ as $\ell \to \infty$. In J, let $\pi/\ell \equiv \Delta\omega$. Then

$$\begin{aligned} J &= \frac{1}{\pi}\sum_{n=1}^{\infty}\left[\int_{-\ell}^{\ell} f(\xi)\cos n\Delta\omega(\xi - x)\, d\xi\right]\Delta\omega \\ &\to \frac{1}{\pi}\int_0^{\infty}\left[\int_{-\infty}^{\infty} f(\xi)\cos\omega(\xi - x)\, d\xi\right] d\omega \end{aligned} \tag{5}$$

as $\ell \to \infty$. To understand the last step, observe that the sum in (5) is a Riemann sum, and as $\Delta\omega \to 0$ (i.e., as $\ell \to \infty$, because $\Delta\omega = \pi/\ell$) it yields a Riemann integral. That is, we partition the interval $0 < \omega < \infty$ into equal parts, of dimension $\Delta\omega = \pi/\ell$, and call $\omega_1 = \pi/\ell$, $\omega_2 = 2\pi/\ell$, $\omega_3 = 3\pi/\ell$, and so on, and use the general Riemann integral formula

$$\lim_{\Delta\omega\to 0}\sum_{n=1}^{\infty} F(n\Delta\omega)\Delta\omega = \lim_{\Delta\omega\to 0}\sum_{n=1}^{\infty} F(\omega_n)\Delta\omega = \int_0^{\infty} F(\omega)\, d\omega. \tag{6}$$

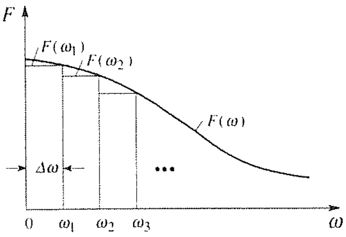


Figure 2. Riemann sum.

(See Fig. 2.) Finally, expressing $\cos\omega(\xi - x) = \cos\omega\xi\cos\omega x + \sin\omega\xi\sin\omega x$, we have

$$\begin{aligned} \text{FS}\, f &= I + J \\ &\to 0 + \int_0^{\infty}\left[\frac{1}{\pi}\int_{-\infty}^{\infty} f(\xi)\cos\omega\xi\, d\xi\right]\cos\omega x d\omega \\ &\quad + \int_0^{\infty}\left[\frac{1}{\pi}\int_{-\infty}^{\infty} f(\xi)\sin\omega\xi\, d\xi\right]\sin\omega x d\omega \\ &= \int_0^{\infty}\left[a(\omega)\cos\omega x + b(\omega)\sin\omega x\right] d\omega \end{aligned} \tag{7}$$

as $\ell \to \infty$. That is, as $\ell \to \infty$ the Fourier series tends to the Fourier integral. We reiterate that our approach has been heuristic, not rigorous. ■

EXAMPLE 1. *Rectangular Pulse.* Let f be the rectangular pulse shown in Fig. 3. This f does satisfy the conditions of the theorem. According to (2b),

$$a(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x\, dx = \frac{1}{\pi}\int_{-1}^{1}\cos\omega x\, dx = \frac{2}{\pi}\frac{\sin\omega}{\omega} \tag{8a}$$

and

$$b(\omega) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x\, dx = 0 \tag{8b}$$

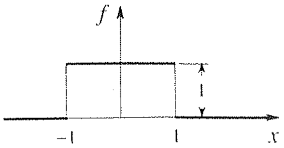


Figure 3. Rectangular pulse.

because the integrand $f(x)\sin\omega x$ is an odd function (recall that even × odd = odd). Thus, the Fourier integral representation of f is, from (2) and (8),

$$f(x) = \frac{2}{\pi}\int_0^{\infty}\frac{\sin\omega}{\omega}\cos\omega x\, d\omega. \tag{9}$$

Just as it was illuminating to plot the *partial sums* $s_n(x)$ of Fourier series, it should be illuminating to see how the *partial integral*

$$f_\Omega(x) = \frac{2}{\pi}\int_0^\Omega \frac{\sin\omega}{\omega}\cos\omega x\, d\omega \tag{10}$$

converges to $f(x)$ as $\Omega \to \infty$. Though the latter integral is not elementary, it can be evaluated in terms of the **sine integral** function

$$\boxed{Si(x) = \int_0^x \frac{\sin t}{t}\, dt,} \tag{11}$$

either analytically (Exercise 1) or using computer software. Using *Maple*, for instance, the command

```
int((2/Pi) * (1/y) * sin (y) * cos (y * x),  y = 0..z);
```

(where z is used in place of Ω) gives

$$f_\Omega(x) = \frac{1}{\pi}\left\{Si[\Omega(x+1)] - Si[\Omega(x-1)]\right\}, \tag{12}$$

which we've plotted in Fig. 4 for $\Omega = 4, 16,$ and 128. Evidently $f_\Omega(x)$ does converge to the pulse $f(x)$, but the convergence is slow near the jump discontinuities; the limiting case as $\Omega \to \infty$ is left for Exercise 3. As in the case of Fourier series, the Fourier integral exhibits the Gibbs phenomenon at such discontinuities.

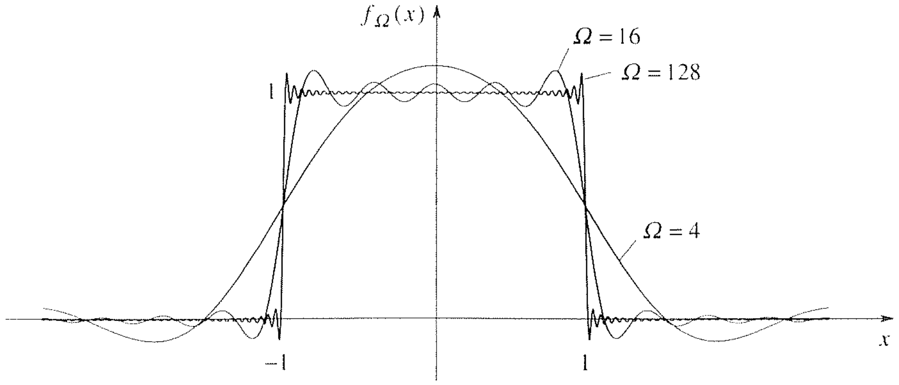


Figure 4. Convergence of $f_\Omega(x)$ to $f(x)$ as $\Omega \to \infty$.

It is also interesting to plot the Fourier coefficient $a(\omega)$, given by (8a), since $a(\omega)$ tells us the harmonic content of f, that is, the amplitude or "amount" of each $\cos\omega x$ harmonic present (Fig. 5). ▮

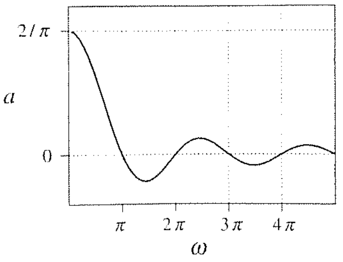


Figure 5. Harmonic content, $a(\omega)$.

EXAMPLE 2. *Infinite Beam on Elastic Foundation.* As a physical application of the Fourier integral let us consider the same problem as contained in Example 4 of Section 17.3 (which we urge you to review), but instead of the periodic loading suppose we have the nonperiodic rectangular pulse loading shown here in Fig. 6. Recall that the beam's

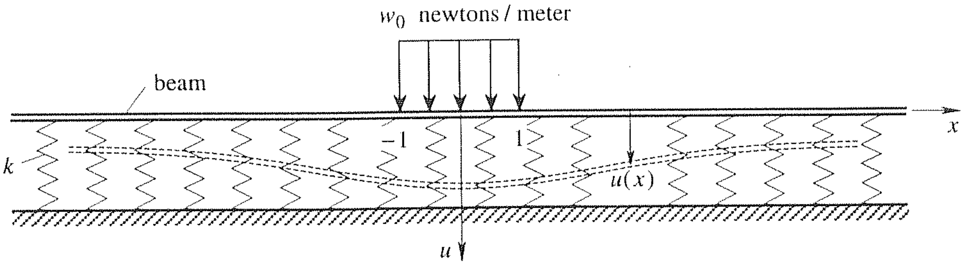


Figure 6. Infinite beam on elastic foundation.

deflection $u(x)$ is governed by the differential equation

$$EIu'''' + ku = w(x), \tag{13}$$

where E, I, k are physical constants, and

$$w(x) = \begin{cases} w_0, & |x| < 1 \\ 0, & |x| > 1 \end{cases} \tag{14}$$

is the rectangular pulse applied load distribution. Proceeding essentially as in Example 4 of Section 17.3, we first express $w(x)$ in Fourier integral form. Since $w(x)$ is merely $f(x)$ in Example 1, scaled by w_0, we conclude from (9) that

$$w(x) = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \cos\omega x \, d\omega. \tag{15}$$

Next, we seek $u(x)$ in the Fourier integral form

$$u(x) = \int_0^\infty a(\omega) \cos\omega x \, d\omega, \tag{16}$$

where $a(\omega)$ remains to be determined. We have omitted the $b(\omega)\sin\omega x$ since $u(x)$ will evidently be a symmetric (even) function of x; that is, if we did include that term we would find that $b(\omega) = 0$.

Formally differentiating (16) under the integral sign four times, and putting that result and (15) and (16) into (13) gives

$$\int_0^\infty (EI\omega^4 + k)a(\omega) \cos\omega x \, d\omega = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \cos\omega x \, d\omega. \tag{17}$$

Then, formally equating the coefficients of each cosine harmonic gives

$$(EI\omega^4 + k)a(\omega) = \frac{2w_0}{\pi} \frac{\sin\omega}{\omega} \tag{18}$$

or

$$a(\omega) = \frac{2w_0}{\pi} \frac{\sin\omega}{\omega} \frac{1}{(EI\omega^4 + k)}, \tag{19}$$

so that (16) becomes

$$u(x) = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \frac{\cos\omega x}{(EI\omega^4 + k)} \, d\omega. \tag{20}$$

This integral can be evaluated analytically using complex variable techniques (the residue theorem), but for our present purposes it will suffice to let (20) stand as it is. ▮

Closure. We obtain the Fourier integral representation of nonperiodic functions defined on $-\infty < x < \infty$ by taking the limit of the Fourier series formula as the period tends to infinity. We limit discussion to two examples because we plan to use the Fourier integral only as a stepping stone to the *Fourier transform*. The latter is more highly developed as a methodology, like the Laplace transform, and is the subject of the next section.

EXERCISES 17.9

1. Use (11) to derive (12) from (10).

2. Derive the Fourier integral representations of the following functions. At which points, if any, does the Fourier integral fail to converge to $f(x)$? To what value does the integral converge at those points?

(a) $f(x) = \begin{cases} 100, & 0 \le x \le 2 \\ 0, & x < 0,\ x > 2 \end{cases}$

(b) $f(x) = \begin{cases} x, & 0 \le x < L \\ 0, & x < 0,\ x \ge L \end{cases}$

(c) $f(x) = \begin{cases} x, & |x| \le L \\ 0, & |x| > L \end{cases}$

(d) $f(x) = \begin{cases} -x, & -5 < x \le 0 \\ 0, & x \le -5,\ x > 0 \end{cases}$

(e) $f(x) = \begin{cases} |x|, & -1 < x < 2 \\ 0, & x \le -1,\ x \ge 2 \end{cases}$

(f) $f(x) = \begin{cases} 10, & 0 \le x \le 3 \\ 5, & 6 \le x \le 9 \\ 0, & x < 0,\ 3 < x < 6,\ x > 9 \end{cases}$

(g) $f(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$

(h) $f(x) = \begin{cases} e^{x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$

(i) $f(x) = e^{-|x|}$

3. (a) Show that $Si(x)$ is an odd function of x.
(b) Using the known integral

$$Si(\infty) = \int_0^\infty \frac{\sin t}{t}\, dt = \frac{\pi}{2}$$

and recalling (11), show that

$$\lim_{\Omega\to\infty} f_\Omega(x) = \begin{cases} 1, & |x| < 1 \\ 1/2, & |x| = 1 \\ 0, & |x| > 1 \end{cases}$$

so that the Fourier integral (9) does converge to the rectangular pulse and (in accordance with Theorem 17.9.1) to the average values at the two jump discontinuities.

4. Comparing (2) with the classical Fourier series (1), it might appear that $a(0)$ is analogous to the a_0 term in the Fourier series and represents the average value of f. Is it true that $a(0)$ is the average value of f where, by the average value of f we mean $\lim_{A\to\infty} \frac{1}{2A}\int_{-A}^{A} f(x)\, dx$? Explain.

17.10 Fourier Transform

Our purpose in this section is to recast the Fourier integral representation of a function f as a pair of formulas, the first giving the so-called Fourier transform of f, and the second giving the inverse of that transform. Once the transform and its inverse are derived, in Section 17.10.1, discussion will closely parallel our discussion, in Chapter 5, of the Laplace transform.

17.10.1. Transition from Fourier integral to Fourier transform. Our starting point is the Fourier integral formula

$$f(x) = \int_0^\infty [a(\omega)\cos\omega x + b(\omega)\sin\omega x]\, d\omega, \tag{1a}$$

where

$$\begin{aligned} a(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x\, dx, \\ b(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x\, dx. \end{aligned} \tag{1b}$$

Just as one can express a Fourier series in complex exponential form (Section 17.3.4), one can express the Fourier integral (1) in complex exponential form. To obtain that form put (1b) into (1a). [First we change the dummy integration variable x in (1b) to ξ, say, to avoid confusing that variable with the x's occurring in (1a), which denote the fixed point at which $f(x)$ is being computed.] Thus,

$$\begin{aligned} f(x) &= \frac{1}{\pi}\int_0^\infty \left\{ \int_{-\infty}^{\infty} f(\xi)[\cos\omega\xi\cos\omega x + \sin\omega\xi\sin\omega x]\, d\xi \right\} d\omega \\ &= \frac{1}{\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)\cos\omega(\xi - x)\, d\xi\, d\omega, \end{aligned} \tag{2}$$

since $\cos(A-B) = \cos A\cos B + \sin A\sin B$. To introduce complex exponentials, re-express (2) as

$$\begin{aligned} f(x) &= \frac{1}{\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)\frac{e^{i\omega(\xi-x)} + e^{-i\omega(\xi-x)}}{2}\, d\xi\, d\omega \\ &= \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{i\omega(\xi-x)}\, d\xi\, d\omega + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\, d\omega. \end{aligned}$$

To combine the two terms on the right-hand side, let us change the dummy integration variable from ω to $-\omega$ in the first. Thus,

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_0^{-\infty} \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\,(-d\omega) \\ &\quad + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\, d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^0 \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\, d\omega + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\, d\omega, \end{aligned}$$

so

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_{-\infty}^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\, d\xi\, d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^\infty \left[\int_{-\infty}^{\infty} f(\xi)e^{-i\omega\xi}\, d\xi\right] e^{i\omega x}\, d\omega. \end{aligned} \tag{3}$$

The latter can be split apart as

$$f(x) = a\int_{-\infty}^{\infty} c(\omega)e^{i\omega x}\,d\omega \tag{4a}$$

and

$$c(\omega) = b\int_{-\infty}^{\infty} f(\xi)e^{-i\omega\xi}\,d\xi, \tag{4b}$$

if the constants a, b are such that $ab = 1/2\pi$. We can make (4) resemble (1) more closely by choosing $a = 1$ and $b = 1/2\pi$, but we will choose $a = 1/2\pi$ and $b = 1$.* There is no longer a need to distinguish x and ξ, because the x's are confined to (4a) and the ξ's to (4b). Thus, to minimize nomenclature and to mimic the form of (1), we write

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} c(\omega)e^{i\omega x}\,d\omega, \tag{5a}$$

$$c(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx. \tag{5b}$$

Rather than thinking of (5a) as the Fourier integral of f and (5b) as giving (give or take the factor of $1/2\pi$) the Fourier coefficients $c(\omega)$, we can think of (5a,b) as a transform pair: (5b) defines the **Fourier transform** $c(\omega)$ of the given function $f(x)$, and (5a) is called the **inversion formula** because putting $c(\omega)$ in and integrating gives us back $f(x)$. It is standard to use the notation $\hat{f}(\omega)$, in place of $c(\omega)$, for the transform, so we rewrite (5) in final form as

$$\boxed{F\{f(x)\} = \hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx,} \tag{6a}$$

and

$$\boxed{F^{-1}\{\hat{f}(\omega)\} = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}\,d\omega.} \tag{6b}$$

Thus, the Fourier transform and inversion formulas are not mysterious; together, they simply amount to the Fourier integral representation, expressed in complex exponential form, and conditions imposed on f are the same as in Theorem 17.9.1.

Let us illustrate the calculation of the transform $\hat{f}(\omega)$ of $f(x)$.

EXAMPLE 1. *Rectangular Pulse.* Consider the rectangular pulse $f(x) = H(x+1) - H(x-1)$, where H denotes the Heaviside function. The graph of f is given in Fig. 1. Using (6a), the Fourier transform of f is

$$\begin{aligned}\hat{f}(\omega) &= \int_{-\infty}^{\infty} [H(x+1) - H(x-1)]e^{-i\omega x}\,dx = \int_{-1}^{1} e^{-i\omega x}\,dx \\ &= \frac{e^{-i\omega x}}{-i\omega}\bigg|_{-1}^{1} = 2\frac{\sin\omega}{\omega}.\end{aligned} \tag{7}$$

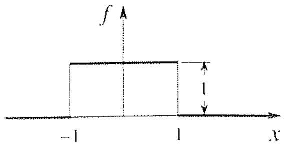


Figure 1. Rectangular pulse.

*Some authors, perhaps out of a greater sense of fair play, choose $a = b = 1/\sqrt{2\pi}$.

COMMENT. We could illustrate the inversion formula as well, by putting $\hat{f}(\omega) = (2\sin\omega)/\omega$ into the integrand of (6b), integrating, and showing that the result is the rectangular pulse f that we started with. In fact, that integral can be evaluated by using the residue theorem of the complex integral calculus, but we won't study that theorem until Chapter 24. ▮

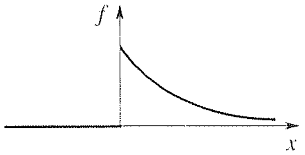


Figure 2. $f(x) = H(x)e^{-ax}$.

EXAMPLE 2. Evaluate the Fourier transform of

$$f(x) = H(x)e^{-ax}, \qquad (a > 0) \tag{8}$$

the graph of which is given in Fig. 2. From (6a),

$$\begin{aligned}\hat{f}(\omega) &= \int_{-\infty}^{\infty} H(x)e^{-ax}e^{-i\omega x}\,dx = \int_0^{\infty} e^{-(a+i\omega)x}\,dx \\ &= -\frac{e^{-(a+i\omega)x}}{a+i\omega}\bigg|_0^{\infty} = \frac{1}{a+i\omega} - \frac{1}{a+i\omega}\lim_{x\to\infty} e^{-(a+i\omega)x} \\ &= \frac{1}{a+i\omega}.\end{aligned} \tag{9}$$

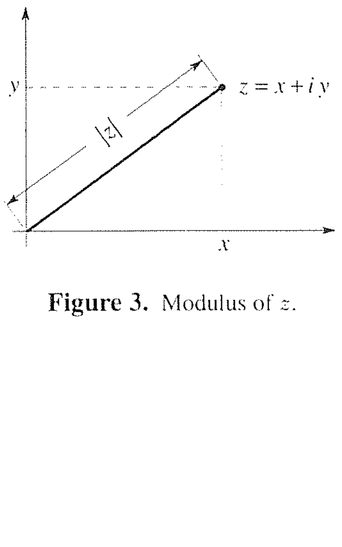


Figure 3. Modulus of z.

To explain the last step in (9), recall that the magnitude (or modulus) $|z|$ of a complex number $z = x + iy$ (Fig. 3) is $|z| = \sqrt{x^2 + y^2}$, and that the modulus of a product is the product of the moduli ($|z_1 z_2| = |z_1|\,|z_2|$). Then

$$\begin{aligned}\left|e^{-(a+i\omega)x}\right| &= \left|e^{-ax}e^{-i\omega x}\right| = \left|e^{-ax}\right|\left|e^{-i\omega x}\right| \\ &= e^{-ax}\left|\cos\omega x - i\sin\omega x\right| \\ &= e^{-ax}\sqrt{\cos^2\omega x + \sin^2\omega x} \\ &= e^{-ax} \to 0\end{aligned} \tag{10}$$

as $x \to \infty$. ▮

As with the Laplace transform, it is convenient to use tables for both the transform and its inverse, insofar as possible. The table provided here, in Appendix D, is brief, but will suffice for present purposes. Much more extensive tables are available,* as well as powerful computer software; the relevant commands, using *Maple*, are given at the end of this section.

17.10.2. Properties and applications. The Fourier transform admits a number of useful properties, our discussion of which will closely parallel our analogous discussion for the Laplace transform, given in Chapter 5. We will assume, without reiteration, that the functions being transformed satisfy the conditions given in Theorem 17.9.1.

*See, for example, A. Erdèlyi (ed.), *Tables of Integral transforms*, Vols. 1 and 2 (New York: McGraw-Hill, 1954). Volume 1 contains Fourier, Laplace, and Mellin transforms, and Volume 2 contains Hankel and various other transforms.

Linearity of the transform and its inverse. For any scalars α and β, and any functions f and g,

$$\boxed{F\{\alpha f + \beta g\} = \alpha F\{f\} + \beta F\{g\}} \tag{11}$$

and

$$\boxed{F^{-1}\{\alpha \hat{f} + \beta \hat{g}\} = \alpha F^{-1}\{\hat{f}\} + \beta F^{-1}\{\hat{g}\}.} \tag{12}$$

Of course, $F^{-1}\{\hat{f}\}$ is f and $F^{-1}\{\hat{g}\}$ is g. Proofs of (11) and (12) follow the same lines as the proofs of Theorems 5.3.1 and 5.3.2, respectively.

Transform of nth derivative. If $f(x), f'(x), \ldots, f^{(n-1)}(x)$ all tend to zero as $x \to \pm\infty$, and $\int_{-\infty}^{\infty} |f^{(j)}(x)|\, dx$ converges for each $j = 0, 1, \ldots, n$, then

$$\boxed{F\{f^{(n)}(x)\} = (i\omega)^n \hat{f}(\omega). \qquad (n = 0, 1, 2, \ldots)} \tag{13}$$

Proof: Let us prove (13) by induction. First, observe that (13) holds for $n = 0$, by definition. To complete the proof by induction we need to establish that if it holds for $n = k$ then it also holds for $n = k + 1$. To do so, integrate by parts:

$$\begin{aligned} F\{f^{(k+1)}(x)\} &= \int_{-\infty}^{\infty} f^{(k+1)}(x) e^{-i\omega x}\, dx \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + i\omega \int_{-\infty}^{\infty} f^{(k)}(x) e^{-i\omega x}\, dx \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + i\omega[(i\omega)^k \hat{f}(\omega)] \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + (i\omega)^{k+1} \hat{f}(\omega), \end{aligned} \tag{14}$$

the third equality following from the assumption that (13) holds for $n = k$. Now, $|f^{(k)}(x) e^{-i\omega x}| = |f^{(k)}(x)|\,|e^{-i\omega x}| = |f^{(k)}(x)|$, and since $f^{(k)}(x) \to 0$ as $x \to \pm\infty$, by assumption, the boundary term in (14) drops out. Thus $F\{f^{(k+1)}(x)\} = (i\omega)^{k+1} \hat{f}(\omega)$ so (13) does hold for $n = k + 1$, which result completes our proof by induction. ■

Fourier convolution. We denote the **Fourier convolution** of functions f and g as $f * g$. It too is a function of x, defined as

$$\boxed{(f * g)(x) \equiv \int_{-\infty}^{\infty} f(x - \xi) g(\xi)\, d\xi.} \tag{15}$$

Then the **Fourier convolution theorem** states that

$$\boxed{F\{f * g\} = \hat{f}(\omega) \hat{g}(\omega),} \tag{16}$$

or, equivalently, that

$$\boxed{F^{-1}\{\hat{f}\hat{g}\} = f * g.} \tag{17}$$

Proof: It suffices to prove (16) or (17), because of their equivalence. Let us prove (16). We have

$$\begin{aligned} F\{f * g\} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x-\xi)g(\xi)\,d\xi \right\} e^{-i\omega x}\,dx \\ &= \int_{-\infty}^{\infty} g(\xi)\,d\xi \int_{-\infty}^{\infty} f(x-\xi)e^{-i\omega x}\,dx \\ &= \int_{-\infty}^{\infty} g(\xi)\,d\xi \int_{-\infty}^{\infty} f(\eta)e^{-i\omega(\xi+\eta)}\,d\eta \\ &= \left(\int_{-\infty}^{\infty} g(\xi)e^{-i\omega\xi}\,d\xi \right) \left(\int_{-\infty}^{\infty} f(\eta)e^{-i\omega\eta}\,d\eta \right) \\ &= \hat{g}(\omega)\hat{f}(\omega), \end{aligned} \tag{18}$$

as was to be shown. In the second equality we reversed the order of integration,* and the third equality follows from the change of variables $x - \xi = \eta$. ■

It is easy to show that the Fourier convolution is commutative,

$$f * g = g * f, \tag{19}$$

proof of which is left for the exercises. That is, it doesn't matter whether we take the argument of f to be $x - \xi$ and the argument of g to be ξ, in (15), or visa versa.

Translation formulas, x-shift and ω-shift. It is also easy to derive the shift (or translation) formulas

$$\boxed{F\{f(x-a)\} = e^{-ia\omega}\hat{f}(\omega)} \tag{20}$$

and

$$\boxed{F^{-1}\{\hat{f}(\omega - a)\} = e^{iax}f(x),} \tag{21}$$

proofs of which are left for the exercises.

Each of the properties (11), (12), (13), (16), (17), (20), and (21) corresponds to an analogous property of the Laplace transform. Properties (11) and (12) are identical to the corresponding properties of the Laplace transform. The derivative property (13) is similar to the property

$$L\{f^{(n)}(t)\} = s^n\overline{f}(s) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0), \tag{22}$$

*Sufficient conditions for the validity of this reversal are that f and g both be **absolutely integrable**, i.e., that $\int_{-\infty}^{\infty} |f(x)|\,dx$ and $\int_{-\infty}^{\infty} |g(x)|\,dx$ both converge. But these conditions have already been assumed. For detailed discussion, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957), p.491.

but does not include any boundary values analogous to the initial conditions $f(0), \ldots, f^{(n-1)}(0)$ in (22), because the boundaries are at $x = \pm\infty$, and it is assumed that $f(x), f'(x), \ldots, f^{(n-1)}(x)$ all tend to zero as $x \to \pm\infty$. The convolution properties (16) and (17) are identical to the corresponding ones for the Laplace transform, but keep in mind that the Fourier and Laplace convolutions are not quite the same; in contrast with the Fourier convolution (15), the Laplace convolution was defined as

$$(f * g)(t) \equiv \int_0^t f(t-\tau)g(\tau)\,d\tau. \tag{23}$$

That is, the integration limits are different.

In the examples to follow, we illustrate the use of these various properties and the table in Appendix D.

EXAMPLE 3. Given

$$f(x) = 4e^{-|x|} - 5e^{-3|x+2|}, \tag{24}$$

evaluate its transform $\hat{f}(\omega)$. First, the linearity property (11) gives

$$F\{4e^{-|x|} - 5e^{-3|x+2|}\} = 4F\{e^{-|x|}\} - 5F\{e^{-3|x+2|}\}. \tag{25}$$

Next, entry 4 in Appendix D gives

$$F\{e^{-|x|}\} = \frac{2}{\omega^2+1}, \tag{26}$$

and entry 11 (with $a = 3$ and $b = 6$) gives

$$\begin{aligned} F\{e^{-3|x+2|}\} &= F\{e^{-|3x+6|}\} \\ &= \frac{1}{3}e^{i6\omega/3}\left(F\{e^{-|x|}\}\right)\Big|_{\omega\to\omega/3} \\ &= \frac{1}{3}e^{i2\omega}\left(\frac{2}{\omega^2+1}\right)\Big|_{\omega\to\omega/3} \\ &= \frac{1}{3}e^{i2\omega}\frac{2}{\left(\frac{\omega}{3}\right)^2+1}. \end{aligned} \tag{27}$$

From (25)–(27), it follows that

$$\begin{aligned} \hat{f}(\omega) &= 4\frac{2}{\omega^2+1} - 5\left(\frac{2}{3}\right)\frac{e^{i2\omega}}{\left(\frac{\omega}{3}\right)^2+1} \\ &= \frac{8}{\omega^2+1} - \frac{30e^{i2\omega}}{\omega^2+9}. \quad \blacksquare \end{aligned}$$

EXAMPLE 4. Given

$$f(x) = xe^{-4x^2}, \tag{28}$$

evaluate $\hat{f}(\omega)$. First,

$$\begin{aligned} F\{e^{-4x^2}\} &= F\{e^{-(2x)^2}\} \\ &= \frac{1}{2}\left(F\{e^{-x^2}\}\right)\Big|_{\omega\to\omega/2} \qquad \text{(entry 11, } a = 2,\ b = 0) \\ &= \frac{1}{2}\sqrt{\pi}\, e^{-\omega^2/4}\Big|_{\omega\to\omega/2} \qquad \text{(entry 5)} \\ &= \frac{\sqrt{\pi}}{2}\, e^{-\omega^2/16}. \end{aligned} \tag{29}$$

Next, entry ~~16~~ 17 (with $n = 1$) gives

$$\hat{f}(\omega) = i\,\frac{d}{d\omega}\left(\frac{\sqrt{\pi}}{2}\, e^{-\omega^2/16}\right) = -i\,\frac{\sqrt{\pi}}{16}\,\omega e^{-\omega^2/16}. \quad ▮$$

EXAMPLE 5. Given $\hat{f}(\omega) = e^{-2|\omega|}$, evaluate $f(x)$. With $a = 2$, entry 1 gives

$$F^{-1}\left\{\frac{\pi}{2}\, e^{-2|\omega|}\right\} = \frac{1}{x^2+4}. \tag{30}$$

Then it follows from the linearity property (12), with $\alpha = \pi/2$ and $\beta = 0$, that (30) can be re-expressed as

$$\frac{\pi}{2}F^{-1}\left\{e^{-2|\omega|}\right\} = \frac{1}{x^2+4}, \tag{31}$$

so

$$F^{-1}\left\{e^{-2|\omega|}\right\} = \frac{2}{\pi}\,\frac{1}{x^2+4}, \tag{32}$$

is the desired inverse function $f(x)$. ▮

EXAMPLE 6. Evaluate

$$F^{-1}\left\{\frac{1}{\omega^2+4\omega+13}\right\}. \tag{33}$$

First, completing the square, write $\omega^2 + 4\omega = (\omega^2 + 4\omega + 4) - 4 = (\omega + 2)^2 - 4$, so (33) becomes

$$F^{-1}\left\{\frac{1}{(\omega+2)^2+9}\right\}. \tag{34}$$

Now, entry 4 (with $a = 3$) and linearity give

$$F^{-1}\left\{\frac{1}{\omega^2+9}\right\} = \frac{1}{6}\, e^{-3|x|}. \tag{35}$$

Finally, from entry 12, with $a = 1$ and $b = 2$, and (35), we have

$$F^{-1}\left\{\frac{1}{(\omega+2)^2+9}\right\} = e^{-i2x}\left(\frac{1}{6}\, e^{-3|x|}\right) \tag{36}$$

as the desired result.

COMMENT. Alternatively, we could have used partial fractions, as follows. Solving $\omega^2 + 4\omega + 13 = 0$ gives $\omega = -2 \pm 3i$. Thus,

$$\begin{aligned}\frac{1}{\omega^2 + 4\omega + 13} &= \frac{1}{(\omega + 2 - 3i)(\omega + 2 + 3i)}\\ &= \frac{1}{6i}\frac{1}{\omega + 2 - 3i} - \frac{1}{6i}\frac{1}{\omega + 2 + 3i}\\ &= \frac{1}{6}\frac{1}{i\omega + (3 + 2i)} - \frac{1}{6}\frac{1}{i\omega + (-3 + 2i)}\\ &= \frac{1}{6}\frac{1}{i\omega + (3 + 2i)} + \frac{1}{6}\frac{1}{-i\omega + (3 - 2i)}.\end{aligned} \tag{37}$$

Observe that the first term after the third equal sign can be inverted by entry 2, because $\text{Re}\,(3 + 2i) = 3 > 0$,* but the second cannot, because $\text{Re}\,(-3 + 2i) = -3 < 0$. Thus, in the final step we rearranged that term so that it can be inverted by entry 3, because $\text{Re}\,(3 - 2i) = 3 > 0$. Finally, using the linearity property (12), together with entries 2 and 3 gives

$$\begin{aligned}F^{-1}\left\{\frac{1}{\omega^2 + 4\omega + 13}\right\} &= \frac{1}{6}H(x)e^{-(3+2i)x} + \frac{1}{6}H(-x)e^{(3-2i)x}\\ &= \frac{1}{6}e^{-i2x}[H(x)e^{-3x} + H(-x)e^{3x}]\\ &= \frac{1}{6}\,e^{-i2x}e^{-3|x|},\end{aligned} \tag{38}$$

as obtained in (36). ▮

EXAMPLE 7. Evaluate

$$F^{-1}\left\{\frac{1}{\omega^2 + 1}\right\}. \tag{39}$$

The latter can be evaluated directly from entry 4 (together with the linearity property), as $e^{-|x|}/2$. However, for pedagogical purposes let us use this example to illustrate the convolution property. First, factor $1/(\omega^2 + 1)$ as

$$\frac{1}{\omega^2 + 1} = \frac{1}{(\omega + i)(\omega - i)} = \frac{1}{(1 - i\omega)(1 + i\omega)}, \tag{40}$$

the final form in (40) being more convenient than the intermediate form because each factor can be inverted, using entries 3 and 2, respectively:

$$F^{-1}\left\{\frac{1}{1 - i\omega}\right\} = H(-x)e^{x}, \tag{41a}$$

$$F^{-1}\left\{\frac{1}{1 + i\omega}\right\} = H(x)e^{-x}. \tag{41b}$$

*If $z = x + iy$ is a complex number, then x and y are called the **real** and **imaginary parts** of z, and are denoted as Re z and Im z, respectively.

Then the Fourier convolution formula (17) gives

$$
\begin{aligned}
F^{-1}\left\{\frac{1}{\omega^2+1}\right\} &= [H(-x)e^{x}] * [H(x)e^{-x}] \\
&= \int_{-\infty}^{\infty} H(-x+\xi)e^{x-\xi}H(\xi)e^{-\xi}\,d\xi \\
&= \begin{cases} e^{x}\displaystyle\int_{x}^{\infty} e^{-2\xi}\,d\xi, & x>0 \\ e^{x}\displaystyle\int_{0}^{\infty} e^{-2\xi}\,d\xi, & x<0 \end{cases} \\
&= \begin{cases} \dfrac{1}{2}e^{-x} & x>0 \\ \dfrac{1}{2}e^{x} & x<0 \end{cases} \\
&= \frac{1}{2}e^{-|x|}. \qquad (42)
\end{aligned}
$$

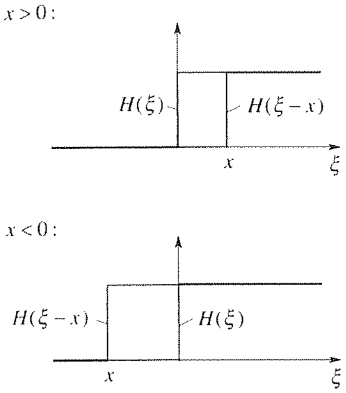


Figure 4. $H(\xi - x)$ and $H(\xi)$.

To understand the third equality in (42), it is useful to plot $H(\xi - x)$ and $H(\xi)$, both for $x > 0$ and for $x < 0$, as we have in Fig. 4. Namely, we see from Fig. 4 that their product $H(\xi - x)H(\xi)$ is $H(\xi - x)$ if $x > 0$, and $H(\xi)$ if $x < 0$.

COMMENT. Alternatively, we could have applied partial fractions to the right-hand side of (40) and proceeded as in the comment in Example 6. ∎

EXAMPLE 8. *Infinite Beam on Elastic Foundation, Revisited.* In Example 2 in Section 17.9 we investigated the deflection $u(x)$ of an infinitely long beam resting on an elastic foundation and subjected to a load $w(x)$ newtons per meter (Fig. 5), governed by the differential equation

$$EIu'''' + ku = w(x), \tag{43}$$

where E, I, and k are physical constants. There we used the Fourier integral method, whereby $w(x)$ was expanded in a Fourier integral, and $u(x)$ was sought in the form of a Fourier integral. Here, we use the Fourier transform instead.

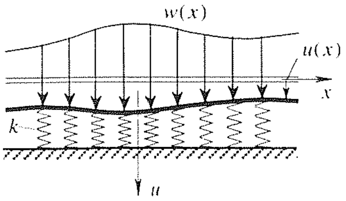


Figure 5. Beam on elastic foundation.

Taking the Fourier transform of (43) (i.e., multiplying each term by $e^{-i\omega x}\,dx$ and integrating from $x = -\infty$ to $x = +\infty$), we have

$$F\{EIu'''' + ku\} = F\{w\}. \tag{44}$$

It follows that

$$EI\,F\{u''''\} + k\,F\{u\} = F\{w\} \tag{45}$$

by the linearity property (11), and

$$EI(i\omega)^4\hat{u} + k\hat{u} = \hat{w} \tag{46}$$

by (13), assuming that u, u', u'', u''' all tend to zero as $x \to \pm\infty$ and that $\int_{-\infty}^{\infty} |u^{(j)}(x)|\,dx$ converges for $j = 0, 1, 2, 3, 4$. Now solving (46) for $\hat{u}$,

$$\hat{u} = \frac{\hat{w}}{EI\omega^4 + k}. \tag{47}$$

To invert (47), let us write the right-hand side in the more suggestive product form

$$\hat{u} = \left(\frac{1}{EI\omega^4 + k} \right) (\hat{w}). \tag{48}$$

The inverse of the product is not the product of the two inverses, but is the Fourier convolution of the two inverses,

$$u(x) = F^{-1} \left\{ \frac{1}{EI\omega^4 + k} \right\} * F^{-1} \{\hat{w}\}. \tag{49}$$

From entry 8 in Appendix D, and the linearity of F^{-1}, we obtain

$$F^{-1} \left\{ \frac{1}{EI\omega^4 + k} \right\} = \frac{\alpha}{\sqrt{2}\, k} e^{-\alpha|x|} \sin(\alpha|x| + \frac{\pi}{4}), \tag{50}$$

where $\alpha = [k/(4EI)]^{1/4}$, and of course $F^{-1}\{\hat{w}\} = w(x)$, so (49) gives

$$u(x) = \frac{\alpha}{\sqrt{2}\, k} \int_{-\infty}^{\infty} e^{-\alpha|x-\xi|} \sin(\alpha|x - \xi| + \frac{\pi}{4}) w(\xi)\, d\xi \tag{51}$$

as the desired solution.

COMMENT 1. It is always important to check our results. An excellent partial check of a result is provided by any special case for which the exact solution is known. In the present example, such a special case is provided by the case where $w(x) = \text{constant} \equiv W$, for then surely the solution $u(x)$ will be a constant too. Specifically, with $w(x) = W$, and u a constant, (43) gives $u(x) = W/k$. If we set $w(\xi) = W$ in (51), and integrate, do we obtain the same result? Yes, but we leave that calculation for the exercises.

In fact, observe that the correctness of (51) for the case where $w(x)$ is a constant is a surprise since some of the assumptions that were built into our solution are violated in that case. For instance, the transform $\hat{w}$ of $w(x) = W$ does not even exist, because the transform integral does not converge. That is,

$$\begin{aligned} \hat{w} = \int_{-\infty}^{\infty} W e^{-i\omega x}\, dx &= W \lim_{A\to\infty} \int_{-A}^{A} e^{-i\omega x}\, dx \\ &= \frac{2W}{\omega} \lim_{A\to\infty} \sin \omega A, \end{aligned} \tag{52}$$

which limit does not exist. Without elaborating, we state that (51) ends up being correct, for this case, even though (44) is not (since $F\{w\} = \hat{w}$ does not exist), thanks to the interchange in the order of integration that underlies the convolution property.* In any case, the moral is that it is often best to proceed formally to a solution. If we can verify that the solution thus obtained does satisfy all of the specified requirements (such as a differential equation and boundary conditions), then there is no need to worry about any lack of rigor in the intermediate steps.

COMMENT 2. If w is a delta function, $w(x) = \delta(x)$, then $u(x)$ is the response to a unit load at $x = 0$ (Fig. 6).† Remember that

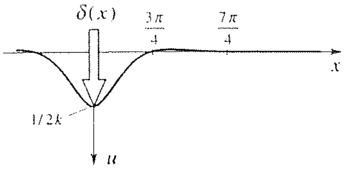


Figure 6. Response to a point unit load at $x = 0$.

*See our proof of the convolution formula, following (17).

†A *unit* load because $\int_{-\infty}^{\infty} w(x)\, dx = \int_{-\infty}^{\infty} \delta(x)\, dx = 1$.

$$\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(0) \tag{53}$$

if f is continuous, by the definition of the delta function. Thus, if $w(x) = \delta(x)$ in (51) then we obtain the response

$$u(x) = \frac{\alpha}{\sqrt{2}\,k} e^{-\alpha|x|} \sin\left(\alpha|x| + \frac{\pi}{4}\right), \tag{54}$$

which is plotted in Fig. 6 for the representative case where $\alpha = 1$.

COMMENT 3. Let us re-express (51) in the compact form

$$u(x) = \int_{-\infty}^{\infty} K(x-\xi)w(\xi)\,d\xi. \tag{55}$$

That is, the output function u is given by the action of an integral operator on the input function w, the operator being multiplication by K followed by integration from $-\infty$ to ∞. We have written $K(x-\xi)$ rather than $K(\xi, x)$ to show that K depends only on the difference between x and ξ. We call K the **kernel** of the integral operator.[*] Since K is a function of the difference $x - \xi$, we call it a **difference kernel**. The physical significance of K is revealed by the fact that if the loading $w(\xi)$ in (51) is a point unit load $\delta(\xi)$ at the origin, then the resulting deflection $u(x)$ is, according to (53),

$$u(x) = \int_{-\infty}^{\infty} K(x-\xi)\delta(\xi)\,d\xi = K(x). \tag{56}$$

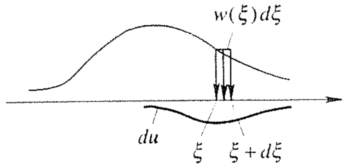


Figure 7. Incremental deflection du, due to load $w(\xi)\,d\xi$.

That is, $K(x)$ is the response function (54), the graph of which is shown in Fig. 6 (for $\alpha = 1$), so $K(x-\xi)$ in (55) is the deflection due to a point unit load at ξ. If the load between ξ and $\xi + d\xi$ is $w(\xi)\,d\xi$ rather than unity (Fig. 7), then the contribution du to the deflection u, due to that bit of loading, is $K(x-\xi)$ scaled by $w(\xi)\,d\xi$,

$$du = K(x-\xi)w(\xi)\,d\xi. \tag{57}$$

Adding all of these du's gives the integral in (55). Thus, we can now understand (55) as a *superposition* principle, whereby $u(x)$ is the sum, or superposition, of the individual contributions du due to each point load $w(\xi)\,d\xi$. That superposition result is a consequence of the linearity of (43). ▮

Additional physical ODE (ordinary differential equation) applications, to the deflection of a loaded string and to the steady-state concentration distribution in a stream subjected to an input of pollutant are given in the exercises. We urge you to at least read through those exercises even if you do not work them.

Closure. We have seen that the Fourier transform and the corresponding inversion formula are actually just a restatement of the Fourier integral representation of a nonperiodic function defined on $-\infty < x < \infty$. We have also seen that the

[*] To further illustrate the idea of the kernel of an integral operator, we note that the kernel of the Laplace transform operator is e^{-st}.

Fourier transform methodology is analogous, and in some aspects identical, to the Laplace transform methodology studied in Chapter 5. Given that similarity, the important question arises, as to which of the two transform methods to use in a given ODE application. The general guideline is as follows:

- The **Laplace transform** is tailored to **initial-value problems** on a **semi-infinite interval** $0 < t < \infty$.
- The **Fourier transform** is tailored to **boundary-value problems** on an **infinite interval** $-\infty < x < \infty$.

The independent variables are usually t (time) and x (linear dimension), but not always. Further, there does exist a "bilateral" Laplace transform defined on an infinite interval, the Laplace transform can sometimes be used for boundary-value problems on finite intervals, and the Fourier transform can (and will, in Section 17.11) be adapted to boundary-value problems on a semi-infinite interval $0 < x < \infty$. But the guideline given above provides the general rule of thumb for selecting one transform or the other.

In closing, we note that to compute the Fourier transform $\hat{f}(\omega)$ of $f(x)$ requires an integration of f over $(-\infty, \infty)$. One can evaluate the integral numerically by sampling the integrand at discrete x points and obtain what is known as the **discrete Fourier transform**. Further, there exist algorithms known collectively as the **fast Fourier transform** (i.e., the **FFT**) that provide a more efficient method of calculating the sum in the discrete Fourier transform. For an introduction to these methods we refer the interested reader to Peter V. O'Neil's *Advanced Engineering Mathematics*, 3rd ed. (Belmont, CA: Wadsworth, 1991).

Computer software. To obtain the Fourier transform of $e^{-5|x|}$, say, using *Maple*, enter

readlib(fourier):

to access the Fourier transform and inverse transform commands. Then enter

$$\text{fourier}(\exp(-5 * \text{abs}(x)),\ x,\ w);$$

and return. The result is

$$10\,\frac{1}{25+w^2}$$

which agrees with entry 4 in Appendix D. To invert the latter, enter

$$\text{invfourier}(10/(25+w\hat{}2),\ w,\ x);$$

and return. The result is

$$e^{-5x}\,\text{Heaviside}(x) + e^{5x}\,\text{Heaviside}(-x)$$

which does reduce to $e^{-5|x|}$.

EXERCISES 17.10

1. Using (6a), derive the result

$$F\{H(x)e^{-ax}\} = \frac{1}{a+i\omega}$$

if Re $a > 0$. (This case was worked in Example 2, but there a was considered to be real. Here, allow for a to be complex, with Re $a > 0$.)

2. Using (6a), derive the result

$$F\{H(-x)e^{ax}\} = \frac{1}{a-i\omega}$$

if Re $a > 0$.

3. Using (6a), derive the result

$$F\{e^{-a|x|}\} = \frac{2a}{\omega^2+a^2}$$

if $a > 0$.

4. Given $\hat{f}(\omega)$, use (6b) to evaluate the inverse, $f(x)$.

(a) $e^{-a|\omega|} \quad (a > 0)$
(b) $H(\omega+a) - H(\omega-a) \quad (a > 0)$
(c) $[H(\omega) - H(\omega-1)]\omega$
(d) $H(\omega)e^{-a\omega} \quad (\text{Re}\, a > 0)$
(e) $[H(\omega+1) - H(\omega-1)]|\omega|$
(f) $\delta(\omega - a)$

5. Derive the following entry in Appendix D, by showing that the transform of the function given in the f column is the result given in the $\hat{f}$ column.

(a) Entry 10 (b) Entry 11
(c) Entry 12 (d) Entry 13
(e) Entry 14 (f) Entry 15
(g) Entry 17, by formally differentiating both sides of

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx$$

a sufficient number of times, with respect to ω
(h) Entry 19, using integration by parts

6. Evaluate the following using Appendix D. You may need to use more than one entry. Cite, by number, any entries that you use.

(a) $F\left\{4x^2e^{-3|x|}\right\}$ (b) $F\left\{xe^{-x^2}\right\}$
(c) $F\left\{\dfrac{\cos 3x}{x^2+2}\right\}$ (d) $F\left\{\dfrac{\sin 2x}{4x^2+3}\right\}$
(e) $F\left\{\dfrac{3}{2x^2+1} - 5e^{-|x|}\right\}$ (f) $F\left\{e^{-|x|} + e^{-3|x+2|}\right\}$
(g) $F^{-1}\left\{\dfrac{4\sin\omega}{\omega} - \dfrac{1}{\sqrt{|\omega|}}\right\}$ (h) $F^{-1}\left\{e^{-2|\omega-3|}\right\}$
(i) $F^{-1}\left\{\dfrac{9}{2\omega+i}\right\}$ (j) $F^{-1}\left\{e^{-\omega^2+4\omega}\right\}$
(k) $F^{-1}\left\{e^{-|\omega|}\cos\omega\right\}$ (l) $F^{-1}\left\{\omega e^{-3\omega^2}\right\}$
(m) $F^{-1}\left\{\dfrac{1}{\omega^2+i\omega+2}\right\}$ (n) $F^{-1}\left\{\dfrac{\omega}{\omega^2+1}\right\}$

7. (a)–(n) Use computer software to obtain (if possible) the transform or inverse transform called for in the corresponding part of Exercise 6.

8. We claimed, in Comment 1 in Example 8, that if $w(x) = W$ then the integral on the right-hand side of (51) is W/k. Verify that claim by evaluating the integral. HINT: Break the integral into two parts, one from $-\infty$ to x and the other from x to ∞, then make a suitable change of variables in each.

9. (*Preservation of evenness and oddness*)

(a) Show that $f(x)$ is an even function of x if and only if $\hat{f}(\omega)$ is an even function of ω.
(b) Show that $f(x)$ is an odd function of x if and only if $\hat{f}(\omega)$ is an odd function of ω.

10. (*Extension of transform tables*) It follows from the results in Exercise 9 that the transform of the even and odd parts of $f(x)$ are the even and odd parts of $\hat{f}(\omega)$, respectively. This result can be used to obtain more information from a given transform table. For example, consider entry 2 in Appendix D. Breaking f and $\hat{f}$ into even and odd parts, show that

$$F\left\{e^{-a|x|}\right\} = \frac{2a}{\omega^2+a^2} \qquad (10.1)$$

and

$$F\left\{(\text{sgn}\, x)e^{-a|x|}\right\} = -\frac{2i\omega}{\omega^2+a^2}, \qquad (10.2)$$

where

$$\operatorname{sgn} x \equiv \begin{cases} +1, & x > 0 \\ -1, & x < 0, \end{cases} \tag{10.3}$$

which is read as "sign of x." Of these two results, observe that (10.1) is identical to entry 4, and that (10.2) is not contained in Appendix D.

11. (*Extension of transform tables*) Another idea that enables us to extend a given Fourier transform table is that of reciprocity, namely, the reciprocity relations

$$F\{\hat{f}(x)\} = 2\pi f(-\omega) \tag{11.1}$$

and

$$F^{-1}\{f(-\omega)\} = \frac{\hat{f}(x)}{2\pi}. \tag{11.2}$$

(a) Derive the relations (11.1) and (11.2).
(b) To illustrate, use (11.1) and entry 4, in Appendix D, to show that

$$F\left\{\frac{2a}{x^2+a^2}\right\} = 2\pi e^{-a|\omega|}, \qquad (a > 0)$$

or, equivalently,

$$F\left\{\frac{1}{x^2+a^2}\right\} = \frac{\pi}{a} e^{-a|\omega|}. \qquad (a > 0)$$

(In this case the result does not extend our table since it already appears as Entry 1.)
(c) Use (11.1) and entry 9 to show that

$$F\left\{\frac{\sin ax}{x}\right\} = \pi[H(\omega+a) - H(\omega-a)]. \qquad (a > 0)$$

(d) Use (11.1) and entry 3 to show that

$$F\left\{\frac{1}{a-ix}\right\} = 2\pi H(\omega) e^{-a\omega}. \qquad (\operatorname{Re} a > 0)$$

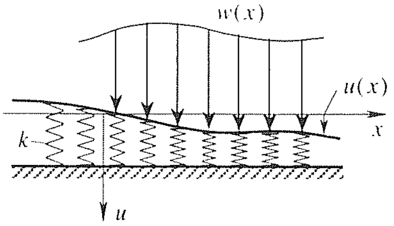
12. (*Deflection of loaded string*) Related to the problem of a beam on an elastic foundation is the analogous problem for a flexible string. Imagine a string (of negligible mass per unit length) stretched along the x axis, over $-\infty < x < \infty$, by a tension T newtons, and let $w(x)$ be an applied load distribution (newtons/meter), as sketched in the figure. If the displacement $u(x)$ of the string is resisted by a distributed spring of stiffness k (newtons per meter per meter), and the slope $u'(x)$ is sufficiently small over $-\infty < x < \infty$, then $u(x)$ is accurately governed by the differential equation $Tu'' - ku = -w$, or,

$$u'' - \alpha^2 u = -f(x), \qquad (-\infty < x < \infty) \tag{12.1}$$

where $\alpha^2 = k/T$ and $f(x) = w(x)/T$. (Here we consider the static deflection. In Chapter 20 we will derive the governing differential equation for the not-necessarily-static case.) Assume that $w(x)$ is sufficiently localized for $u(x)$ to satisfy the boundary conditions

$$u \to 0 \quad \text{and} \quad u' \to 0 \tag{12.2}$$

as $x \to \pm\infty$, as well.



(a) Solving (12.1) and (12.2) using a Fourier transform, derive the solution

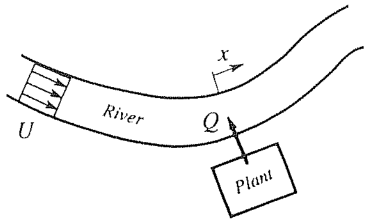
$$u(x) = -\frac{1}{2\alpha}\int_{-\infty}^{\infty} e^{-\alpha|x-\xi|} f(\xi)\, d\xi. \tag{12.3}$$

(b) Evaluate $u(x)$ from (12.3) for the case where $f(x)$ is a point unit load at the origin, $f(x) = \delta(x)$, and sketch the graph of u.
(c) Verify, by formally using the Leibniz rule, that (12.3) satisfies the differential equation (12.1).

13. (*Pollution in river*) Suppose that a manufacturing plant discharges a certain pollutant into an initially clear river at the rate Q grams/second. We wish to determine the resulting steady-state distribution of pollutant in the river, i.e., its concentration c (grams/meter^3). Measure x along the river, positive downstream, with origin at the plant site, as shown in the figure. The river flows with velocity U (meters/second), and has a cross-sectional area A, both of which, for simplicity, we assume to be constant. Also for simplicity, suppose that c is a function of x only; i.e., it is a constant over each cross section of the stream. This is evidently a poor approximation near the plant, where we expect appreciable across-stream and vertical variations in c, but it should suffice if we are concerned mostly with the concentration variation far upstream and downstream (for $|x|$ greater than several river widths, say). Then it can be shown that $c(x)$ is governed by the differential equation

$$kc'' - Uc' - \beta c = -\frac{Q}{A}\,\delta(x), \tag{13.1}$$

where k (meters^2/second) is a diffusion constant, β (grams per second per gram) is a chemical decay constant, and $\delta(x)$ is a delta function. [Physically, (13.1) expresses a mass balance between the *input* $Q\delta(x)/A$, the transport of pollutant by *diffusion*, kc'', the transport of pollutant by *convection* with the moving stream, Uc', and by disappearance through *chemical decay*, βc.]



(a) Applying the Fourier transform to (13.1), show that

$$c(x) = F^{-1}\left\{\frac{Q}{A}\,\frac{1}{k\omega^2 + iU\omega + \beta}\right\}. \tag{13.2}$$

(b) Expanding $1/(k\omega^2 + iU\omega + \beta)$ in partial fractions, and then using Appendix D, show that (13.2) gives

$$c(x) = \begin{cases} c_0 e^{-\Omega_- x}, & x < 0 \\ c_0 e^{-\Omega_+ x}, & x > 0 \end{cases} \tag{13.3}$$

where

$$c_0 = \frac{Q}{A\sqrt{U^2 + 4k\beta}},$$

$$\Omega_\pm = \frac{1}{2k}\left(\pm\sqrt{U^2 + 4k\beta} - U\right).$$

(c) Sketch the graph of $c(x)$ and state the qualitative effect of increasing β.

17.11 Fourier Cosine and Sine Transforms, and Passage from Fourier Integral to Laplace Transform (Optional)

17.11.1. Cosine and sine transforms. Recall from Section 17.4 that if a problem is defined on a finite interval, say $0 < x < L$, then the concepts of periodicity and Fourier series are not directly applicable. However, by fictitiously extending both the domain and the functions involved to the infinite interval $-\infty < x < \infty$, so that the extended functions are periodic, we are able to use Fourier series representations of those functions. Depending upon the symmetries and/or antisymmetries about $x = 0$ and $x = L$, we obtain half- or quarter-range cosine or sine expansions.

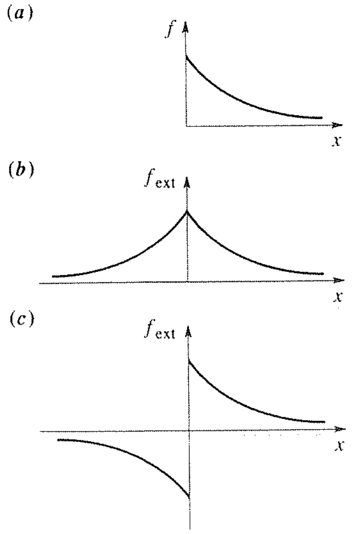


Figure 1. Even and odd extensions of f.

Similarly, we sometimes encounter problems defined on a semi-infinite interval, say $0 < x < \infty$. Fictitiously extending both the domain and the functions involved, to the infinite interval $-\infty < x < \infty$, we will be able to use Fourier integral representations of those functions. Specifically, given a function f defined on $0 < x < \infty$ (such as the one shown in Fig. 1a), we shall be interested in two particular extensions, one that is even and one that is odd, as indicated in Figs.1b and 1c, respectively.

Denoting the extended functions as f_{ext}, consider first the even extension (Fig. 1b). Then the Fourier transform of f_{ext} is

$$\widehat{f_{\text{ext}}}(\omega) = \int_{-\infty}^{\infty} f_{\text{ext}}(x) e^{-i\omega x}\, dx$$

$$= \int_{-\infty}^{\infty} f_{\text{ext}}(x)(\cos \omega x - i \sin \omega x)\, dx = 2 \int_{0}^{\infty} f_{\text{ext}}(x) \cos \omega x \, dx, \quad \text{(1a)}$$

where the last equality holds because $f_{\text{ext}}(x) \cos \omega x$ is an even function of x and $f_{\text{ext}}(x) \sin \omega x$ is odd, and the inversion formula gives

$$\begin{aligned} f_{\text{ext}}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f_{\text{ext}}}(\omega) e^{i\omega x} \, d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f_{\text{ext}}}(\omega)(\cos \omega x + i \sin \omega x)\, d\omega \\ &= \frac{2}{2\pi} \int_{0}^{\infty} \widehat{f_{\text{ext}}}(\omega) \cos \omega x \, d\omega, \end{aligned} \quad \text{(1b)}$$

where the last equality holds because $\widehat{f_{\text{ext}}}(\omega)$ is an even function of ω [as can be seen from (1a)] and $\sin \omega x$ is an odd function of ω, so that the term $\widehat{f_{\text{ext}}}(\omega) \sin \omega x$ in the integrand is odd. Putting (1a) into (1b) gives the single statement

$$f_{\text{ext}}(x) = \int_{0}^{\infty} \left\{ \frac{2}{\pi} \int_{0}^{\infty} f_{\text{ext}}(\xi) \cos \omega\xi \, d\xi \right\} \cos \omega x \, d\omega. \quad (-\infty < x < \infty) \quad (2)$$

Since (2) holds on $-\infty < x < \infty$, it also holds on the original interval $0 < x < \infty$. For $0 < x < \infty$ we can drop the subscripted "ext" on the left-hand side, because on that interval $f_{\text{ext}}(x) = f(x)$. Further, we can drop the "ext" on the right-hand side because the ξ integral is on $0 < x < \infty$, not $-\infty < x < \infty$. Then (2) becomes

$$f(x) = \frac{2}{\pi} \int_{0}^{\infty} \left\{ \int_{0}^{\infty} f(\xi) \cos \omega\xi \, d\xi \right\} \cos \omega x \, d\omega, \qquad (0 < x < \infty) \quad (3)$$

which can be re-expressed, equivalently, as the **Fourier cosine transform**

$$\boxed{F_C\{f(x)\} = \hat{f}_C(\omega) = \int_{0}^{\infty} f(x) \cos \omega x \, dx} \quad \text{(4a)}$$

and its inverse

$$\boxed{F_C^{-1}\{\hat{f}_C(\omega)\} = f(x) = \frac{2}{\pi} \int_{0}^{\infty} \hat{f}_C(\omega) \cos \omega x \, d\omega,} \quad \text{(4b)}$$

on $0 < x < \infty$.

Similarly, an odd extension of f gives the **Fourier sine transform**

$$\boxed{F_S\{f(x)\} = \hat{f}_S(\omega) = \int_{0}^{\infty} f(x) \sin \omega x \, dx} \quad \text{(5a)}$$

and its inverse

$$\boxed{F_S^{-1}\{\hat{f}_S(\omega)\} = f(x) = \frac{2}{\pi} \int_{0}^{\infty} \hat{f}_S(\omega) \sin \omega x \, d\omega,} \quad \text{(5b)}$$

on $0 < x < \infty$ (Exercise 1). Sufficient conditions on f, for (4) and (5) to hold, are that f and f' be piecewise continuous on $0 \leq x < \infty$, and that $\int_0^\infty |f(x)|\, dx$ converge (i.e., that f be *absolutely integrable* on $0 < x < \infty$).

Next, we could derive the various properties of the cosine and sine transforms, that are analogous to those derived for the Fourier transform. However, for brevity let it suffice to state that both the cosine and sine transforms, and their inverses, are linear, and that the transforms of the derivative f' are

$$\boxed{\begin{aligned} F_C\{f'(x)\} &= \omega \hat{f}_S(\omega) - f(0), \\ F_S\{f'(x)\} &= -\omega \hat{f}_C(\omega), \end{aligned}} \tag{6a,b}$$

if we assume additionally that $f(x) \to 0$ as $x \to \infty$. Let us derive (6a) and leave (6b) for the exercises. Integrating by parts,

$$\begin{aligned} F_C\{f'(x)\} &= \int_0^\infty f'(x) \cos \omega x \, dx \\ &= [f(x) \cos \omega x]\Big|_0^\infty + \omega \int_0^\infty f(x) \sin \omega x \, dx \\ &= 0 - f(0) + \omega \hat{f}_S(\omega). \end{aligned} \tag{7}$$

The transforms of higher-order derivatives can be obtained by repeated use of (6). For instance, replacing the function f by the function f' in (6) gives

$$F_C\{f''(x)\} = \omega F_S\{f'(x)\} - f'(0) = \omega[-\omega \hat{f}_C(\omega)] - f'(0) \tag{8a}$$

and

$$F_S\{f''(x)\} = -\omega F_C\{f'(x)\} = -\omega[\omega \hat{f}_S(\omega) - f(0)], \tag{8b}$$

so

$$\boxed{\begin{aligned} F_C\{f''(x)\} &= -\omega^2 \hat{f}_C(\omega) - f'(0), \\ F_S\{f''(x)\} &= -\omega^2 \hat{f}_S(\omega) + \omega f(0), \end{aligned}} \tag{9a,b}$$

if f and f' tend to zero as $x \to \infty$. Similarly for higher-order derivatives. For convolution properties, see Exercise 10 and Appendix E.

Short Fourier cosine and sine transform tables are given in Appendix E.

EXAMPLE 1. Consider the boundary-value problem

$$u'' - 9u = 50e^{-2x}, \qquad (0 < x < \infty) \tag{10a}$$

$$u(0) = u_0, \quad u(\infty) \text{ bounded.} \tag{10b}$$

To solve (10) using an integral transform we need to choose among the Laplace, Fourier cosine, and Fourier sine transforms, all of these being candidates because they are semi-infinite transforms; that is, they apply when the domain is semi-infinite ($0 < x < \infty$ in this case). The Laplace transform will be inconvenient at best, because it is tailored to initial value problems whereas (10) is of boundary-value type. In choosing between the Fourier cosine and sine transforms, the key is in (9). Taking a cosine transform of (10a) we will, according to (9a), need to know $u'(0)$, yet the latter is not prescribed; taking a sine transform of (10a) we will, according to (9a), need to know $u(0)$, and the latter *is* prescribed.

Thus, take the Fourier sine transform of (10a), using the linearity of the transform, property (9b), and entry $1S$ in Appendix E:

$$F_S\{u''-9u\}=F_S\left\{50e^{-2x}\right\}, \tag{11a}$$

$$F_S\{u''\}-9F_S\{u\}=50F_S\left\{e^{-2x}\right\}, \tag{11b}$$

$$-\omega^2\hat{u}_S+\omega u_0-9\hat{u}_S=50\,\frac{\omega}{\omega^2+4}. \tag{11c}$$

Solving this linear algebraic equation for $\hat{u}_S$ gives

$$\hat{u}_S(\omega)=u_0\,\frac{\omega}{\omega^2+9}-50\,\frac{\omega}{(\omega^2+4)(\omega^2+9)}. \tag{12}$$

By partial fractions,

$$\frac{1}{(\omega^2+4)(\omega^2+9)}=\frac{1}{5}\,\frac{1}{\omega^2+4}-\frac{1}{5}\,\frac{1}{\omega^2+9}, \tag{13}$$

so

$$\hat{u}_S(\omega)=(u_0+10)\frac{\omega}{\omega^2+9}-10\frac{\omega}{\omega^2+4}. \tag{14}$$

Then, using the linearity of the inverse transform and entry $1S$, gives

$$\begin{aligned}u(x)&=F_S^{-1}\left\{(u_0+10)\frac{\omega}{\omega^2+9}-10\frac{\omega}{\omega^2+4}\right\}\\&=(u_0+10)F_S^{-1}\left\{\frac{\omega}{\omega^2+9}\right\}-10F_S^{-1}\left\{\frac{\omega}{\omega^2+4}\right\}\\&=(u_0+10)e^{-3x}-10e^{-2x}.\end{aligned} \tag{15}$$

COMMENT 1. Note that we used property (9b) tentatively since (9b) presupposes that both $u(x)\to 0$ and $u'(x)\to 0$ as $x\to\infty$. Now that $u(x)$ is in hand, in (15), we can check, and we see that these conditions are met. Or, more directly, we can simply verify that (15) does satisfy (10a) and (10b).

COMMENT 2. If $u'(0)$ were prescribed, in (10b), in place of $u(0)$, then we would use the Fourier cosine transform instead. ▮

17.11.2. Passage from Fourier integral to Laplace transform. We have seen that the Fourier transform is merely a restatement of the Fourier integral, the Fourier integral is a limiting case of the Fourier series of a periodic function (as the period tends to infinity), and the Fourier series is as an eigenfunction expansion corresponding to a periodic Sturm–Liouville problem. On the other hand, the Laplace transform and its inversion formula were given in Chapter 5 without derivation and may seem unrelated to these Fourier methods. In this final subsection we will show that the Laplace transform and its inverse can be derived from the Fourier integral!

We begin with the Fourier integral in the complex exponential form given by equation (3) in Section 17.10,

$$F(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} F(\tau)e^{-i\omega\tau}\,d\tau\right\}e^{i\omega t}\,d\omega, \qquad (-\infty < t < \infty) \tag{16}$$

where we use t in place of x because we will end up with the Laplace transform, in which the independent variable is traditionally taken to be t (because in applications it usually corresponds to the time). Further, it will be convenient to use F in place of f. In (16), let

$$F(t) = \begin{cases} e^{-\gamma t}f(t), & t > 0 \\ 0, & t < 0, \end{cases} \tag{17}$$

where γ is a real constant which is sufficiently positive so that $e^{-\gamma t}$"clobbers" $f(t)$ as $t \to \infty$. Specifically, suppose that f is of exponential order (defined in section 5.2) as $t \to \infty$. Then a sufficiently positive γ can indeed be found such that $e^{-\gamma t}f(t)$ dies out exponentially fast as $t \to \infty$. Of course, whereas $e^{-\gamma t}$ helps as $t \to +\infty$, it hurts as $t \to -\infty$. Thus, we simply "shut off F" for $t < 0$ by defining it to be zero for $t < 0$, in (17). The resulting F easily satisfies the condition $\int_{-\infty}^{\infty}|F(t)|\,dt < \infty$ contained in the Fourier integral theorem.

Putting (17) into (16) gives

$$H(t)e^{-\gamma t}f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{0}^{\infty} e^{-\gamma\tau}f(\tau)e^{-i\omega\tau}\,d\tau\right\}e^{i\omega t}\,d\omega \tag{18}$$

where H is the Heaviside function. Thus,

$$H(t)f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{0}^{\infty} e^{-(\gamma+i\omega)\tau}f(\tau)\,d\tau\right\}e^{(\gamma+i\omega)t}\,d\omega \tag{19}$$

which form suggests changing variables from ω to s according to

$$s = \gamma + i\omega. \tag{20}$$

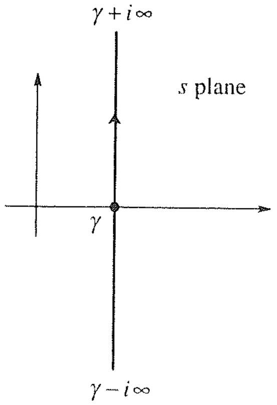


Figure 2. Line of integration in s plane.

Thus,

$$H(t)f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty}\left\{\int_{0}^{\infty} e^{-s\tau}f(\tau)\,d\tau\right\}e^{st}\,ds, \tag{21}$$

where $\int_{\gamma-i\infty}^{\gamma+i\infty}$ denotes an integration along a vertical line in a complex s plane (Fig. 2). If we define the **Laplace transform** of f as

$$\boxed{L\{f(t)\} = \overline{f}(s) = \int_{0}^{\infty} f(t)e^{-st}\,dt,} \tag{22}$$

where we have changed the dummy integration variable from τ to t, then (21) gives the inversion formula as

$$\boxed{L^{-1}\{\overline{f}(s)\} = H(t)f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \overline{f}(s)e^{st}\,ds.} \tag{23}$$

Normally we are interested only in $0 < t < \infty$, so we can replace $H(t)$ by 1 in (23). However, for completeness we observe that while the integral in (23) gives $f(t)$ for $t > 0$, it gives 0 for $t < 0$, which result is, after all, in accordance with our definition $F(t) = 0$ for $t < 0$, in (17).

Closure. In this section we have combined two topics, each being too brief to justify a full section. First we introduced the Fourier cosine and sine transforms, for problems defined on the semi-infinite interval $0 < x < \infty$, which results are analogous to the half-range cosine and sine Fourier series. The Laplace transform is also a semi-infinite transform, but it is tailored to *initial-value* problems whereas the Fourier cosine and sine transforms are tailored to *boundary-value* problems, because $L\{f^{(n)}(t)\}$ involves values of $f, f', \dots, f^{(n-1)}$ at $t = 0$, whereas $F_C\{f^{(n)}(x)\}$ and $F_S\{f^{(n)}(x)\}$ involve values both at $x = 0$ and at $x = \infty$. Whether to use a cosine transform or a sine transform, in solving a differential equation boundary-value problem, depends (as we saw in Example 1) on the type of boundary conditions prescribed at $x = 0$.

Finally, we used the complex exponential form of the Fourier integral to derive the Laplace transform and its inversion formula. Normally, in using the Laplace transform, we have the t domain of interest is $t > 0$. However, we noted that if one *were* to obtain an inverse Laplace transform by evaluating the inversion integral in (23), for $t < 0$, then one would inevitably find that the inverse function is identically zero for $t < 0$.

EXERCISES 17.11

1. Derive the formulas (5a) and (5b) for the Fourier sine transform and its inverse, respectively.

2. Derive (6b), that $F_S\{f'(x)\} = -\omega \hat{f}_C(\omega)$.

3. Derive these results:

(a) $F_C\{f''''(x)\} = \omega^4\hat{f}_C(\omega) + \omega^2 f'(0) - f'''(0)$

(b) $F_S\{f''''(x)\} = \omega^4\hat{f}_S(\omega) - \omega^3 f(0) + \omega f''(0)$ if $f(x), f'(x), f''(x)$, and $f'''(x)$ all tend to zero as $x \to \infty$.

4. Given the rectangular pulse $f(x) = 50[1 - H(x-4)]$, evaluate $\hat{f}_C(\omega)$ and $\hat{f}_S(\omega)$.

5. Use computer software, such as the int command on *Maple*, to evaluate each, where $a > 0$. Show that your result agrees with the corresponding result obtained from Appendix E.

(a) $F_C\{e^{-ax}\}$ (b) $F_S\{e^{-ax}\}$
(c) $F_C\{xe^{-ax}\}$ (d) $F_S\{xe^{-ax}\}$
(e) $F_C\{x^2e^{-ax}\}$ (f) $F_S\{x^2e^{-ax}\}$

NOTE: Be careful, in using tables or software, to be clear on the author's definition of the transform. For instance, in contrast with our definitions (4) and (5) of the Fourier cosine and sine transforms, some authors use the more symmetric versions

$$F_C\{f(x)\} = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\cos\omega x\,dx \tag{5.1}$$

$$F_C^{-1}\{\hat{f}_C(\omega)\} = \sqrt{\frac{2}{\pi}}\int_0^\infty \hat{f}_C(\omega)\cos\omega x\,d\omega \tag{5.2}$$

and

$$F_S\{f(x)\} = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\sin\omega x\,dx \tag{5.3}$$

$$F_S^{-1}\{\hat{f}_S(\omega)\} = \sqrt{\frac{2}{\pi}}\int_0^\infty \hat{f}_S(\omega)\sin\omega x\,d\omega. \tag{5.4}$$

6. In Example 1 we avoided the Laplace transform because (10) is of boundary-value type, not of initial-value type. Show that the Laplace transform can be used nevertheless, and does give the solution (15), though not as conveniently. HINT: When you take the transform of u'' you will be faced with a $u'(0)$ term, which is not prescribed in (10b). Thus, call that quantity C, say, and evaluate it by imposing on your solution the condition that $u(\infty)$ be bounded, at the end.

7. We used a sine transform to solve Example 1. Try to solve (10) using a cosine transform instead, and explain why that method does not work.

8. Modify (10) by changing $u(0) = u_0$ to $u'(0) = u_0'$, and solve by a cosine or sine transform.

9. Solve, using a cosine or sine transform.

(a) $u'' - 9u = 50e^{-3x}, \quad (0 < x < \infty)$
$u(0) = 0, \; u(\infty)$ bounded

(b) $u'' - 9u = 50e^{-3x}, \quad (0 < x < \infty)$
$u'(0) = 0, \; u(\infty)$ bounded

10. (*Convolution Theorem*) As for the Fourier and Laplace transforms there are convolution theorems for the Fourier cosine and sine transforms, and these are given in Appendix E.

(a) Prove the Fourier cosine transform convolution theorem (entry 7C), either by showing that the transform of the given integral is $\hat{f}_C(\omega)\hat{g}_S(\omega)$ or by showing that the inverse of $\hat{f}_C(\omega)\hat{g}_S(\omega)$ is the given integral.

(b) Prove the Fourier sine transform convolution theorem (entry 7S).

(c) Verify entry 7C for the case where $f(x) = e^{-x}$ and $g(x) = e^{-3x}$.

(d) Verify entry 7S for the case where $f(x) = e^{-x}$ and $g(x) = e^{-3x}$.

Chapter 17 Review

We began with the Fourier series representation

$$f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{\ell} + b_n\sin\frac{n\pi x}{\ell}\right) \qquad (-\infty < x < \infty) \tag{1}$$

of any 2ℓ-periodic function f defined on $-\infty < x < \infty$, which representation is valid subject to very mild conditions on f. For instance, f can even have jump discontinuities, whereas to represent f by a Taylor series f needs to be infinitely differentiable over the interval under consideration (and even that condition does not quite suffice). In applications, periodic functions arise in a number of ways. For instance an offshore structure is probably subjected to wave forces that are periodic in time, and the temperature distribution around the edge of a circular disk is a 2π-periodic function of the polar angle θ.

If, instead, the x domain is finite, say $0 < x < L$, then Fourier series can still be employed – by fictitiously extending the domain to $-\infty < x < \infty$, and extending the definition of f onto that domain so that f_{ext} is periodic. Such extensions can be accomplished in an infinite number of ways, but the four that will be needed, in applications, correspond to extensions that are symmetric or antisymmetric about the ends $x = 0$ and $x = L$, and these are the half- and quarter-range cosine and sine expansions, which we denote by HRC, HRS, QRC, and QRS, respectively:

$$\text{HRC:} \qquad f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L} \qquad (0 < x < L) \tag{2}$$

$$\text{HRS:} \qquad f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L} \qquad (0 < x < L) \tag{3}$$

$$\text{QRC:} \qquad f(x) = \sum_{n=1,3,\dots}^{\infty} a_n \cos\frac{n\pi x}{2L} \qquad (0 < x < L) \tag{4}$$

$$\text{QRS:} \qquad f(x) = \sum_{n=1,3,\dots}^{\infty} b_n \sin\frac{n\pi x}{2L} \qquad (0 < x < L) \tag{5}$$

(For brevity, we do not repeat, here, the formulas for the a_n's and b_n's given in the text.) We emphasized that the choice, as to which of these four expansions to use, will be dictated by the context.

From a vector space viewpoint, (1)–(5) amount to expansions of f in terms of an infinite set of orthogonal base vectors. In (1), for instance, the base vectors are $1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \dots$. Such sets of orthogonal base vectors arise as the eigenfunctions of Sturm–Liouville problems, namely, eigenvalue problems of the type

$$(py')' + qy + \lambda w y = 0, \qquad (a < x < b) \tag{6}$$

with homogeneous boundary conditions at $x = a$ and $x = b$, with the inner product

$$\langle u, v \rangle = \int_a^b u(x)v(x)w(x)\,dx \tag{7}$$

with weight function w. For instance, the base vectors in (1) are the eigenfunctions of the Sturm–Liouville problem

$$\begin{aligned} y'' + \lambda y = 0, &\quad (-\ell < x < \ell) \\ y(-\ell) - y(\ell) = 0, &\quad y'(-\ell) - y'(\ell) = 0, \end{aligned} \tag{8}$$

and the base vectors in (2)–(5) are the eigenfunctions of the Sturm–Liouville problems

$$y'' + \lambda y = 0, \qquad (0 < x < L)$$

$$\begin{aligned} &\text{HRC:} \quad y'(0) = 0, \quad y'(L) = 0, \\ &\text{HRS:} \quad y(0) = 0, \quad y(L) = 0, \\ &\text{QRC:} \quad y'(0) = 0, \quad y(L) = 0, \\ &\text{QRS:} \quad y(0) = 0, \quad y'(L) = 0, \end{aligned} \tag{9}$$

respectively. The four cases in (9) are examples of regular Sturm–Liouville systems, while the Sturm–Liouville problem in (8) is of periodic type. Singular Sturm–Liouville problems were discussed in Section 17.8, prominent examples involving the Bessel and Legendre equations.

We showed that if we let $\ell \to \infty$ in (1), then the frequency spectrum $\{n\pi/\ell\}$ becomes a continuous spectrum from 0 to ∞ and we obtain, in place of the Fourier series (1), the Fourier integral

$$f(x) = \int_0^\infty [a(\omega)\cos\omega x + b(\omega)\sin\omega x]\, d\omega. \qquad (-\infty < x < \infty) \tag{10}$$

Expressing (10) in complex exponential form, we obtained the equation pair

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\, dx, \tag{11a}$$

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}\, d\omega, \tag{11b}$$

which are equivalent to (10) and which are the Fourier transform and its inverse, respectively. Finally, we obtained the Fourier cosine and sine transforms from (11), for problems defined on the semi-infinite interval $0 < x < \infty$, these being analogous to the half-range cosine and sine representations of functions defined on $0 < x < L$.

Chapter 18

Diffusion Equation

18.1 Introduction

Chapters 18–20 are about partial differential equations (PDE's). Two of the most important PDE's of mathematical physics have already been encountered in Chapter 16. In Example 3 in Section 16.8, we derived the equation

$$\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t} \tag{1}$$

governing the unsteady diffusion of heat by conduction, where α^2 is a physical constant known as the diffusivity of the material, and $T(x, y, z, t)$ is the temperature field. The latter is known as the **heat equation** but is also called the **diffusion equation** because it governs diffusion process in general. For instance, whereas (1) governs the diffusion of *heat*, the equation

$$D\nabla^2 c = \frac{\partial c}{\partial t} \tag{2}$$

governs the unsteady diffusion of *material* (such as a particular chemical pollutant within a body of water or of an anti-cancer drug within an organ such as the liver), where D, like α^2, is a concentration of the material (i.e., the mass of material per unit volume of medium); (2) is of the same form as (1).

If a steady state is achieved, then $\partial T/\partial t = 0$ and (1) reduces to the **Laplace equation**

$$\nabla^2 T = 0. \tag{3}$$

In Section 16.10 we found that the Laplace equation also governs the velocity potential $\Phi(x, y, z)$ for irrotational incompressible flows; other applications of the Laplace equation are discussed in Chapter 20.

Arguably, equations (1) and (3) are two of the three most prominent PDE's of mathematical physics, the third being the **wave equation** of the form

$$c^2 \nabla^2 u = \frac{\partial^2 u}{\partial t^2} \tag{4}$$

where c^2 is a constant.

The diffusion, wave, and Laplace equations are the subjects of Chapters 18, 19, and 20, respectively. They could be presented in any order, but we have chosen this sequence for pedagogical reasons that will be explained as we proceed.

Our approach in these three chapters is substantially different from our approach in the early chapters on ODE's (ordinary differential equations). For ODE's we proceeded systematically, beginning with first-order equations and then moving on to equations of second order and higher and developing the general theory – which covered existence, uniqueness, and methods of solution. Though our emphasis was on linear equations, we considered nonlinear equations as well. For PDE's, however, our scope is more limited as we focus almost entirely on the diffusion, wave, and Laplace equations and emphasize solution technique. These three are by no means the only PDE's encountered in applications. but they are extremely important, and the solution methods that we develop can be applied to various other (linear) PDE's as well.

18.2 Preliminary Concepts

18.2.1. Definitions. Recall (from Section 1.2) that a differential equation is a **partial differential equation** if it contains partial derivatives of the dependent variable with respect to two or more independent variables. For most applications the independent variables are one or more space variables (Cartesian or non-Cartesian), and possibly the time t, but the dependent variables encountered are much more varied and include temperature, concentration, deflection of a string or membrane or beam, velocity potential, and electric potential, to name just a few. As for ODE's, there may be more than one dependent variable, and we may have a system of PDE's in two or more unknowns. Here, however, we consider only the case of one equation in one unknown.

Also as for ODE's, we define the **order** of a PDE as the order of the highest derivative therein, and we say that a function is a **solution** of a PDE, over a particular domain of the independent variables, if its substitution into the equation reduces that equation to an identity everywhere within that domain. For instance, the PDE

$$u_{xx} - 5u_y + u = xy^2(y - 15) \tag{1}$$

(where subscripts denote partial derivatives) is of second order and admits the solutions

$$u_1(x, y) = 6e^{2x+y} + xy^3 \qquad \text{and} \qquad u_2(x, y) = -e^{3x+2y} + xy^3$$

over the entire x, y plane (and other solutions as well). For instance, putting u_1 into (1) gives

$$24e^{2x+y} - 5(6e^{2x+y} + 3xy^2) + 6e^{2x+y} + xy^3 = xy^2(y - 15),$$

which is indeed an identity for all values of x and y.

It is standard and convenient to use the differential operator notation $L[\ ]$ for PDE's, as we do for ODE's. Accordingly, (1) may be written more compactly as

$$L[u] = f, \tag{2}$$

where

$$L = \frac{\partial^2}{\partial x^2} - 5\frac{\partial}{\partial y} + 1 \tag{3}$$

is the second-order *partial differential operator*, and $f(x, y) = xy^2(y - 15)$. Just as the definition of a function is not complete until we specify the domain on which it acts [e.g., the function $\sin x$ defined on $0 \le x \le \pi$ is not the same as the function $\sin x$ defined on $-\infty < x < 5$], likewise the definition of a differential operator is not complete until we specify the domain of functions on which it acts. We do not "figure out" what the domain is; we specify it. For L given by (3), for instance, we might specify its domain as the set of functions $u(x, y)$ that are defined on the first quadrant ($0 < x < \infty, 0 < y < \infty$) and that are twice differentiable in x and once in y.

We say that a differential operator, be it an ordinary differential operator or a partial differential operator, is **linear** if

$$\boxed{L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]} \tag{4}$$

for any functions u and v in the domain of L and for any constants α and β; otherwise it is **nonlinear**.

EXAMPLE 1. The operator L given by (3) is linear because

$$\begin{aligned} L[\alpha u + \beta v] &= \left(\frac{\partial^2}{\partial x^2} - 5\frac{\partial}{\partial y} + 1\right)(\alpha u + \beta v) \\ &= \alpha(u_{xx} - 5u_y + u) + \beta(v_{xx} - 5v_y + v) \\ &= \alpha L[u] + \beta L[v]. \quad \blacksquare \end{aligned}$$

EXAMPLE 2. The operator defined by

$$L[u] = u_{xx} + uu_y$$

is nonlinear because the difference

$$\begin{aligned} &L[\alpha u + \beta v] - \alpha L[u] - \beta L[v] \\ &\quad = (\alpha u_{xx} + \beta v_{xx}) + (\alpha u + \beta v)(\alpha u_y + \beta v_y) - \alpha(u_{xx} + uu_y) - \beta(v_{xx} + vv_y) \\ &\quad = (\alpha^2 - \alpha)uu_y + (\beta^2 - \beta)vv_y + \alpha\beta(uv_y + u_y v) \end{aligned}$$

is not identically zero. For instance, if $u(x, y) = y$, $v(x, y) = 0$, $\alpha = 2$, and $\beta = 6$, then the right-hand side is $3y$, not zero. ∎

If an operator L is linear, then it follows immediately from (4) that

$$L[\alpha_1 u_1 + \cdots + \alpha_k u_k] = \alpha_1 L[u_1] + \cdots + \alpha_k L[u_k] \tag{5}$$

for any functions $u_1, \dots, u_k$ in the domain of L and for any constants $\alpha_1, \dots, \alpha_k$, for any finite k.

Consider a linear differential equation

$$L[u] = f, \tag{6}$$

where f is a prescribed function of the independent variables. If $f = 0$ then (6) is **homogeneous**, and if $f \neq 0$ then (6) is **nonhomogeneous** with f as a forcing function. Any solution of the full equation $L[u] = f$ is called a **particular solution** of (6). The power of the linearity property (5) is that it enables us to build a more robust solution from a collection of individual solutions by superposition. For suppose that $u_1, \dots, u_k$ are solutions of the homogeneous version $L[u] = 0$ of (6), and that u_p is a particular solution of (6). Then

$$u = C_1 u_1 + \cdots + C_k u_k + u_p \tag{7}$$

is a solution of (6) for any constants $C_1, \dots, C_k$ since

$$\begin{aligned} L[C_1 u_1 + \cdots + C_k u_k + u_p] &= C_1 L[u_1] + \cdots + C_k L[u_k] + L[u_p] \\ &= C_1(0) + \cdots + C_k(0) + f \\ &= f. \end{aligned} \tag{8}$$

We say that (7) is robust in the sense that it contains k arbitrary constants, which are available to help u to satisfy the various boundary conditions that may be prescribed.

However, there is a major difference in the application of the foregoing idea to ODE's and PDE's. If (6) is a linear kth-order ODE and $u_1, \dots, u_k$ are linearly independent solutions of the homogeneous equation $L[u] = 0$, then (7) is a *general* solution of (6). For the PDE's that we study in these chapters we will be able to find even an infinite number of solutions $u_1, u_2, \dots$, yet

$$u = u_p + \sum_{j=1}^{\infty} C_j u_j \tag{9}$$

may fall short of being a general solution of (6). That is, (6) may admit solutions that cannot be expressed in the form (9) for any choice of the C_j's. Though it would be nice to obtain general solutions of our PDE's (and indeed we will for the wave equation in Chapter 19) our objective is more limited than that. Specifically, we will be content to be able to obtain solutions satisfying specific boundary conditions, and it will turn out that the infinite series solutions that we develop will indeed be robust enough for that, even if they fall short of being general solutions.

18.2.2. Second-order linear equations and their classification. We are especially, though not exclusively, interested in linear second-order PDE's of the form[*]

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f, \tag{10}$$

[*] We assume that A, B, C are not all zero for then (10) would not be of second order. Further, we assume that if $A = B = 0$ then $D \neq 0$, for if $A = B = D = 0$ then only y derivatives appear and (10) would more reasonably be regarded as an ODE rather than as a PDE. Similarly, we assume that if $B = C = 0$ then $E \neq 0$.

where $A, \ldots, F, f$ are prescribed functions of x and y and where the 2 is included for subsequent convenience. Often $A, \ldots, F$ are constants, but not always. The independent variables x and y will be Cartesian space variables, or else x will be a Cartesian space variable and y will be the time – in which case we will use t in place of the generic y.

We classify (10) as one of three types, depending on the sign of the discriminant $B^2 - AC$: (10) is

$$\boxed{\begin{array}{ll} \textbf{parabolic} & \text{if } B^2 - AC = 0, \\ \textbf{hyperbolic} & \text{if } B^2 - AC > 0, \\ \textbf{elliptic} & \text{if } B^2 - AC < 0 \end{array}} \tag{11}$$

in the region under consideration. This terminology is by analogy with the general equation of a conic section, $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$, which gives a parabola, a hyperbola, or an ellipse, according to the sign of the discriminant $b^2 - ac$. Just as parabolas, hyperbolas, and ellipses are governed by distinct geometrical theories, so are parabolic, hyperbolic, and elliptic PDE's governed by distinct theories. Prototype examples of these three types are as follows:

1. The **diffusion equation**

$$\alpha^2 u_{xx} = u_t \qquad (\alpha^2 = \text{constant}) \tag{12}$$

is **parabolic** (with $y \to t$) since $B^2 - AC = 0^2 - (\alpha^2)(0) = 0$.

2. The **wave equation**

$$c^2 u_{xx} = u_{tt} \qquad (c^2 = \text{constant}) \tag{13}$$

is **hyperbolic** (with $y \to t$) since $B^2 - AC = 0^2 - (c^2)(-1) = c^2 > 0$.

3. The **Laplace equation**

$$u_{xx} + u_{yy} = 0 \tag{14}$$

is **elliptic** since $B^2 - AC = 0^2 - (1)(1) = -1 < 0$.

Thus, keep in mind in Chapters 18–20 that the diffusion, wave, and Laplace equations are not only of importance in their own right, but that they are also representative of the three equation types – parabolic, hyperbolic, and elliptic, respectively. For instance, the PDE

$$\alpha^2 u_{xx} = u_t + V u_x + H u \tag{15}$$

governs the diffusion of heat in a one-dimensional rod, but differs from the basic diffusion equation (12) by virtue of the Vu_x term (which is due to the rod being in motion with constant speed) and the Hu term (which is due to heat loss from the lateral surface of the rod to the environment). However, (15) is still parabolic, like (12), since $B^2 - AC = 0 - (\alpha^2)(0) = 0$, so its solution should in fundamental ways be similar to the solution of (12) that is discussed in this chapter. Such variations from the basic diffusion, wave, or Laplace equation, are considered in a number of the end-of-section exercises.

EXAMPLE 3. Since $A, \ldots, F, f$ in (10) may be functions of x and y, the discriminant $B^2 - AC$ may be a function of x and y. Thus, besides the possibility $B^2 - AC$ is zero, positive, or negative everywhere in the x, y plane, it is also possible that it is zero, positive, or negative in different parts of the x, y plane. To illustrate, consider the **Tricomi equation**

$$u_{xx} + xu_{yy} = 0, \tag{16}$$

which arises in the study of the two-dimensional steady transonic flow past a body such as a wing. (Transonic means that flight speed is close to the speed of sound.) Then

$$B^2 - AC = 0 - (1)(x) = -x$$

so the Tricomi equation is elliptic in the right half plane $x > 0$ and hyperbolic in the left half plane $x < 0$. Thus the Tricomi equation is a **change-of-type** equation, with solutions that are qualitatively different in the two half planes. For ODE's, an analogous behavior is exhibited by the **Airy equation**

$$y'' + xy = 0, \tag{17}$$

which has oscillatory solutions for $x > 0$ and nonoscillatory solutions for $x < 0$.

The moral of this example is that a given example of equation (10) might be of one type in one region and of another type in another region. Such cases are more difficult and are not studied here. ∎

18.2.3. Diffusion equation and modeling. In Example 3 of Section 16.8 we derive the heat equation $\alpha^2\nabla^2 u = u_t$ by considering a heat balance for an arbitrary control volume and using the divergence theorem. (There we used T for the temperature field, but here we use u, since we would like to have the letter T available for another purpose.) In slight contrast with the arbitrary control volume approach, engineering textbooks usually consider infinitesimal elements in such derivations. Let us present such a derivation for a one-dimensional rod and, at the same time, include the additional effects of translation of the rod and heat loss to the environment from its lateral surface.

Specifically, consider a uniform rod of cross-sectional area A, circumference s, mass density σ (mass per unit volume), thermal conductivity k, and specific heat c, and suppose it is in uniform motion with speed v in the positive x direction. Consider an element of the rod between x and $x + \Delta x$, where the x axis is fixed in space and the rod is moving relative to it. Assume, merely for definiteness and with no loss of generality, that the derivative u_x is positive at x and at $x + \Delta x$. Then, according to the Fourier law of heat conduction (Example 3, Section 16.8), heat enters the element through its right-hand face at the rate $kAu_x|_{x+\Delta x}$ and leaves through its left-hand face at the rate $kAu_x|_x$ (Fig. 1).

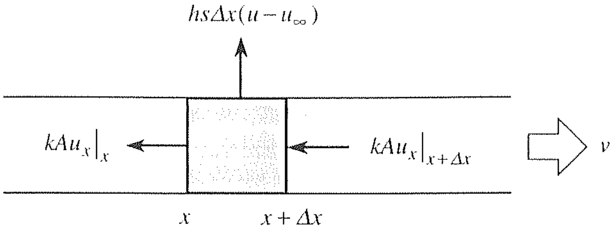


Figure 1. Heat balance for infinitesimal element.

Additionally, suppose there is a heat loss to the environment from the exposed lateral surface of the element, that is proportional to the surface area $s\Delta x$ and the temperature difference $u - u_\infty$, where u_∞ is the temperature of the environment (which we assume to be constant). If the constant of proportionality is some known heat transfer coefficient h, then by *Newton's law of cooling* the rate of heat loss is $hs\Delta x(u - u_\infty)$ (Fig. 1). Finally, there is a transport of heat, in at x and out at $x + \Delta x$, because the rod is translating. Recall that the heat contained in a mass m at (absolute) temperature u is mcu. In time Δt the mass m entering at the left and leaving at the right is the dimension $v\Delta t$ times the area A times the density σ or, on a per unit time basis, $vA\sigma$. Thus, the heat in at the left, per unit time, is $vA\sigma cu|_x$, and the heat out at the right, per unit time, is $vA\sigma cu|_{x+\Delta x}$ (not shown in Fig. 1). Then the net heat influx into the element, per unit time, is

$$kAu_x\Big|_{x+\Delta x} - kAu_x\Big|_{x} - hs\Delta x(u - u_\infty) + vA\sigma cu\Big|_{x} - vA\sigma cu\Big|_{x+\Delta x}. \tag{18}$$

The latter must equal the rate of change of the heat mcu contained in the element, where $m = A\Delta x\sigma$, so

$$kA\left(u_x\Big|_{x+\Delta x} - u_x\Big|_{x}\right) - hs\Delta x(u-u_\infty) - vA\sigma c\left(u\Big|_{x+\Delta x} - u\Big|_{x}\right) = \frac{\partial}{\partial t}\left(A\Delta x\sigma cu\right). \tag{19}$$

Dividing through by $A\Delta x\sigma c$ and letting $\Delta x \to 0$ gives the desired field equation

$$\frac{k}{c\sigma}\,u_{xx} - \frac{hs}{Ac\sigma}\,(u - u_\infty) - v\,u_x = u_t \tag{20}$$

or, more concisely,

$$\alpha^2 u'_{xx} = u'_t + vu'_x + Hu', \tag{21}$$

where $\alpha^2 = k/(c\sigma)$ is the thermal diffusivity of the material, v is the translational speed, $H = hs/(Ac\sigma)$ is proportional to the heat transfer coefficient h, and $u' = u - u_\infty$ is the temperature at any point in the rod *relative* to the ambient temperature u_∞.* If we remember that u' is the relative temperature we can, for notational

*Some common values of α^2 (cm^2/sec) are: silver, 1.70; copper, 1.14; aluminum, 0.86; cast iron, 0.16; brick, 0.0052; glass, 0.0034; and water, 0.0014.

convenience, drop the primes in (21). Of course, (21) is the same as (15), mentioned above.

The derivation given above is only heuristic and is typical of the elemental approach to deriving various field equations, as it is found in engineering science textbooks. But it is not hard to render the derivation rigorous. First, the $A\Delta x\sigma cu$ on the right-hand side of (19) should really have been an integral over the element, but

$$\frac{\partial}{\partial t}\int_x^{x+\Delta x} A\sigma cu(\xi,t)\,d\xi = \frac{\partial}{\partial t}A\sigma cu(x+\mu\Delta x,t)\Delta x \sim A\Delta x\sigma cu_t(x,t) \quad (22)$$

as $\Delta x \to 0$, so we obtain the same result as before. The equality in (22) follows from the mean value theorem of the integral calculus, for which we merely need $u(\xi,t)$ to be a continuous function of x, where μ is some value between 0 and 1. Similarly, the $hs\Delta x(u-u_\infty)$ term in (19) should actually be an integral over the lateral surface, but once again we can use the mean value theorem and obtain the same result as before.

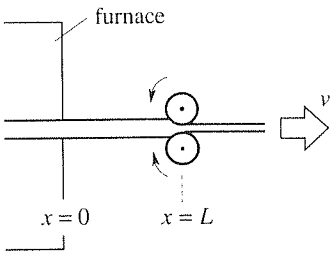


Figure 2. Extrusion.

The Vu_x term in (15) is called a **convection** term, and the Hu term is called a **Newton cooling** term; of course, it will be a heating term rather than a cooling term if the environment is hotter than the rod. As a physical application where both terms would be important, consider a hot continuous metal rod being drawn from a furnace and entering an extrusion die at some distance from the furnace (Fig. 2). In designing the facility it is important to predict the temperature of the rod when it reaches the die, as a function of the various design parameters, so that the metal is still sufficiently hot to be extruded.

Let us set $V=0$ and $H=0$ and limit our subsequent attention to the basic one-dimensional diffusion equation (12). As a typical application of (12), consider the heat conduction through the outer wall of a house (Fig. 3), and let x,y,z be normal, horizontal, and vertical axes, respectively, with $x>L$ corresponding to the interior of the house and $x<0$ corresponding to the exterior. Actually, we should be working with the *three*-dimensional diffusion equation

$$\alpha^2(u_{xx}+u_{yy}+u_{zz}) = u_t. \quad (23)$$

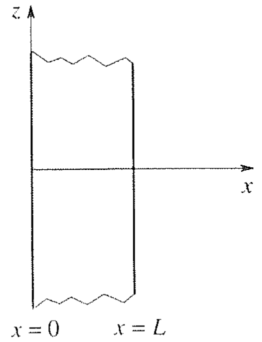


Figure 3. Conduction through a wall.

However, on a cold (or hot) day we expect u to vary very little with y and z compared to its variation with x (over $0<x<L$). Thus, as a reasonable *approximation* we can neglect the u_{yy} and u_{zz} terms, in which case (23) simplifies to the one-dimensional equation

$$\alpha^2 u_{xx} = u_t \quad (24)$$

or, in operator form,

$$L[u] = \left(\alpha^2\frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial t}\right)[u] = \alpha^2 u_{xx} - u_t = 0. \quad (25)$$

To complete the problem formulation it is useful to observe that the domain of interest is the semi-infinite strip $0<x<L$ and $0<t<\infty$ in the x,t plane, as shown in Fig. 4. Clearly, (25) admits an infinite number of solutions such as $4x$, $5x - 30$, $\sin x \exp(-\alpha^2 t)$, $x + 20 - \cos 3x \exp(-9\alpha^2 t)$, and so on. Our expectation, if only intuitive at this stage, is that if we append to (25) a suitable set of boundary conditions, then the resulting problem will have a unique solution. To motivate such a set of conditions, observe that (25) is a second-order equation with respect to x, so we expect two x boundary conditions to be appropriate, one at $x = 0$ and one at $x = L$. If we call the outside temperature u_1, and call the inside temperature u_2, which is the temperature setting on our thermostat, then the boundary conditions on x are

$$u(0,t) = u_1, \qquad (0 < t < \infty) \tag{26a}$$

$$u(L,t) = u_2. \qquad (0 < t < \infty) \tag{26b}$$

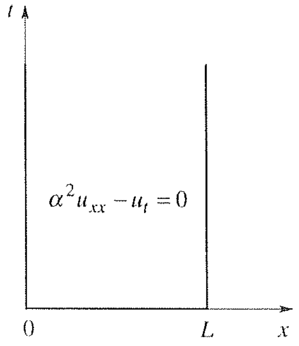


Figure 4. The x, t plane.

Further, (25) is a first-order equation with respect to t, so we expect one t boundary condition to be appropriate, at $t = 0$ (i.e., an initial condition). If we prescribe some initial temperature distribution

$$u(x,0) = f(x), \qquad (0 < x < L) \tag{27}$$

then the full problem statement consists of (25)–(27) and is summarized in Fig. 5. We solve that problem for $u(x,t)$ in Section 18.3, and also establish its uniqueness.

Thus, in formulating the problem the idea is to impose boundary and initial conditions that are sufficient to reduce the solution set to a unique solution, but not excessive so that there is *no* solution. For instance, if we impose not only the conditions shown in Fig. 5 but also the condition $u_x(0,t) = 0$, then that problem will have no solution. Whereas existence and uniqueness questions were prominent in our study of ODE's, here we are not so concerned about the existence question, because we will generally be able to *find* a solution. The nagging question that remains, then, is uniqueness. Thus, representative uniqueness theorems are presented in these chapters.

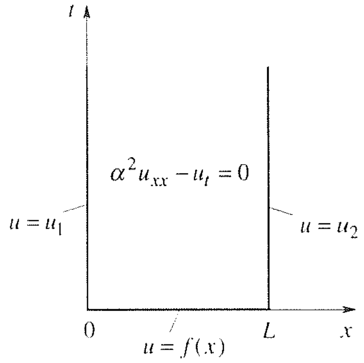


Figure 5. The problem.

Besides thinking of our problem (25)–(27) as governing heat conduction in a wall, we can think of it as governing heat conduction in a rod that is thermally insulated everywhere on its lateral surface – but not at its two ends, which are subjected to temperatures u_1 and u_2 for all $0 < t < \infty$ (Fig. 6).*

Before closing, let us consider the possible boundary conditions at $x = 0$ and $x = L$ more fully. We distinguish three types. Conditions (26a,b) are examples of **boundary conditions of the first kind**, or **Dirichlet boundary conditions**, because they are of the form

$$u \text{ prescribed} \tag{28}$$

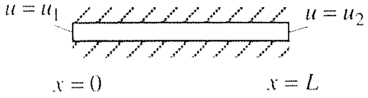


Figure 6. Insulated rod.

on the boundary. In (26a,b) u is prescribed to be a constant, but Dirichlet boundary conditions need not be constant. For instance, if the time of interest is short compared to a day, then it may well suffice to take $u(0,t) = u_1$ to be a constant, but if

*In fact, even for the wall shown in Fig. 3, any "pencil" of material, parallel to the x axis and extending from $x = 0$ to $x = L$, is essentially an insulated rod, insulated by virtue of the lack of temperature variation with y and z.

the time of interest is on the order of a day or longer, then we really need to take $u(0,t) = u_1(t)$ to be a function of time.

A **boundary condition of the second kind**, or **Neumann boundary condition**, is a derivative boundary condition of the form

$$u_n \text{ prescribed,} \tag{29}$$

where u_n denotes the derivative $\partial u/\partial n$ of u normal to the boundary under consideration. In the present example u_n is $-u_x$ on the $x = 0$ boundary and $+u_x$ on the $x = L$ boundary. Physically, (29) amounts to prescribing the *heat flux* rather than the temperature. For instance, the heat flux $Q(t)$ crossing the left end of the rod shown in Fig. 6, counted positive if it flows from left to right is, according to Fourier's law of heat conduction,

$$Q(t) = -kAu_x(0,t). \tag{30}$$

Thus, if we specify the normal derivative $u_x(0,t)$ we are, in effect, specifying the heat flux $Q(t)$. We see from (30) that a homogeneous Neumann boundary condition

$$u_x(0,t) = 0 \tag{31}$$

[or $u_x(L,t) = 0$] amounts to a stipulation that that end is thermally *insulated*, for then $Q(t) = 0$.

Finally, a **boundary condition of the third kind**, or **Robin boundary condition**, occurs when a linear combination of u and u_n is prescribed. To illustrate, within the context of the present example, consider the heat flux crossing the end $x = L$, say. According to the Fourier law of conduction the flux crossing $x = L$ from the left is $-kAu_x(L,t)$, and according to Newton's law of cooling the flux crossing $x = L$ into the environment is $hA[u(L,t) - u_2]$. Since these must be equal we have the boundary condition

$$-kAu_x(L,t) = hA[u(L,t) - u_2] \tag{32}$$

or

$$u_x(L,t) + \frac{h}{k}\,u(L,t) = -\frac{h}{k}\,u_2, \tag{33}$$

which is a boundary condition of Robin type, also called a "mixed" boundary condition. It is useful to nondimensionalize terms in this equation. Nondimensionalizing u with respect to the reference temperature u_2 and x with respect to the reference length L, the nondimensional quantities are

$$\overline{x} = x/L \qquad \text{and} \qquad \overline{u} = u/u_2. \tag{34}$$

Then (32) becomes $-\dfrac{kAu_2}{L}\,\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} = hAu_2\left[\overline{u}\Big|_{\overline{x}=1} - 1\right]$ or

$$-\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} = \text{Bi}\left[\overline{u}\Big|_{\overline{x}=1} - 1\right], \tag{35}$$

where the dimensionless parameter

$$\mathrm{Bi} = \frac{hL}{k} \tag{36}$$

is known in heat transfer as the *Biot number*. If Bi $\gg 1$, then (35) implies that $\overline{u}\Big|_{\overline{x}=1} - 1 \approx 0$ or, returning to dimensional terms,

$$u(L, t) = u_2, \tag{37}$$

which was the boundary condition naively adopted in (26b). However, if the values of h, L, and k give Bi $\ll 1$, then (35) implies that $\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} \approx 0$ or, in dimensional terms,

$$u_x(L, t) = 0. \tag{38}$$

If Bi is neither very large nor very small, then we should leave the mixed boundary condition (33) intact.

With the foregoing remarks completed we will simply specify boundary conditions such as (37), (38), or (33) without further discussion as we now turn our attention to the solution of such problems.

Closure. Keep in mind, as we embark on our study of PDE's in Chapters 18–20, that we are focusing on the extremely important class of PDE's of the form

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f, \tag{39}$$

where $A, \ldots, F, f$ are prescribed functions of x and y. Within that class we distinguish three types. Specifically, we say that (39) is parabolic if $B^2 - AC = 0$, hyperbolic if $B^2 - AC > 0$, and elliptic if $B^2 - AC < 0$. Representative of these types are the diffusion, wave, and Laplace equations, respectively, and we will devote one chapter to each of these types.

Since (39) is of second order, appropriate boundary conditions will involve u and possibly first-order derivatives of u. Specifically, if u, u_n, or a linear combination of u and u_n are prescribed, then we say that the boundary condition is of the first kind (Dirichlet type), second kind (Neumann type), or third kind (Robin type), respectively. Of these, Robin boundary conditions are the most difficult.

EXERCISES 18.2

1. Show that (5) follows from (4). HINT: Use mathematical induction.

2. Show whether the equation is linear or nonlinear; k, α, and β are constants. HINT: See Examples 1 and 2.

(a) Helmholtz equation, $\nabla^2 u + k^2 u = 0$

(b) Korteweg-de Vries (KdV) equation, $u_t + \alpha u u_x + \beta u_{xxx} = 0$

(c) biharmonic equation, $\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0$

(d) Tricomi equation, $u_{xx} + xu_{yy} = 0$

(e) $u_{xx} + u_{yy} = e^u$

(f) $x^4 u_{xx} = u_y$

(g) $u_{xx} + 5u_{xy} - xu = e^x$
(h) $xu_x + uu_y = 6$

3. Classify the following PDE's, defined over $-\infty < x < \infty$ and $-\infty < y < \infty$, as elliptic, parabolic, or hyperbolic. If the equation is of mixed type, identify the relevant regions and give the classification within each region.

(a) $u_{xx} + u_{xy} - x^2 u_y = e^{xy}$
(b) $xu_{xx} - u_{xy} + yu_{yy} + 3u_y = 1$
(c) $u_{xy} + u_x - 4u_y = 6u$
(d) $xu_{xx} - (\sin^2 y + 1)u_{yy} = x^2 u$
(e) $u_{xx} + u_{xy} + u_{yy} + u_x + u_y + u = 1$
(f) $u_{xx} + (\cos x)u_{yy} = 2xy$
(g) $u_{xx} + u_x + u_y = x^3 u$
(h) $u_{xy} - u_{yy} + e^x u = f(x, y)$

4. In deriving the diffusion equation (21), we assumed that the cross-sectional shape of the rod does not vary with x. Reconsider our derivation for the basic case where there is no Newton cooling (i.e., the lateral surface is insulated, $h = 0$) and no translation of the rod ($v = 0$), but allow for the cross-sectional area A to vary with x. Show that the revised diffusion equation is

$$\frac{\alpha^2}{A(x)}[A(x)u_x]_x = u_t. \tag{4.1}$$

18.3 Separation of Variables

18.3.1. The method of separation of variables. We explain the method of separation of variables by a sequence of examples.

EXAMPLE 1. Consider the diffusion problem

$$L[u] = \alpha^2 u_{xx} - u_t = 0, \qquad (0 < x < L,\ 0 < t < \infty) \tag{1a}$$
$$u(0, t) = u_1, \quad u(L, t) = u_2, \qquad (0 < t < \infty) \tag{1b}$$
$$u(x, 0) = f(x), \qquad (0 < x < L) \tag{1c}$$

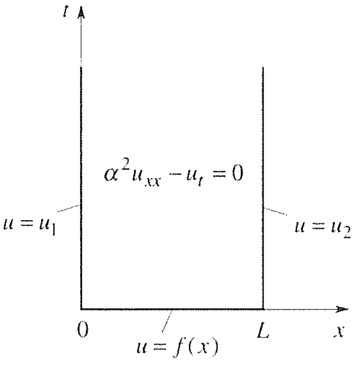


Figure 1. The problem (1).

that is derived in Section 18.2 and that governs the temperature field $u(x, t)$ in a rod with insulated lateral surface (or in a wall or slab of thickness L); see Fig. 1.

According to the method of **separation of variables** we begin by seeking solutions of (1a) in the product form

$$\boxed{u(x, t) = X(x)T(t).} \tag{2}$$

Putting (2) into (1a) gives

$$\alpha^2 X''T = XT', \tag{3}$$

where primes denote ordinary differentiation. To separate the variables, divide both sides of (3) by XT and obtain

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T}. \tag{4}$$

Actually, we divided by α^2 as well, but whether we have $1/\alpha^2$ on the right-hand side of (4) or α^2 on the left-hand side will not affect the final result. Observe that (4) is of the form

$$F(x) = G(t). \tag{5}$$

Since x and t are independent variables, $G(t)$ does not vary with x. But $F(x) = G(t)$, so $F(x)$ does not vary with x either. Hence $F(x)$ is a constant. From (5), $G(t)$ is a constant also, the same constant.* Thus,

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T} = \text{constant} = -\kappa^2, \tag{6}$$

say, where we have written κ^2 for convenience because we will soon need to take the square root of that quantity. Motivation for including the minus sign in (6) is given in the end-of-example comments.

The beauty of the separation procedure is that in place of the *partial* differential equation $\alpha^2 u_{xx} = u_t$ we now have two *ordinary* differential equations,

$$\frac{X''}{X} = -\kappa^2 \qquad \text{and} \qquad \frac{1}{\alpha^2}\frac{T'}{T} = -\kappa^2,$$

or,

$$X'' + \kappa^2 X = 0, \tag{7a}$$

$$T' + \kappa^2\alpha^2 T = 0. \tag{7b}$$

The *separation constant* κ remains to be determined. Solving (7a,b) gives

$$X = A\cos\kappa x + B\sin\kappa x, \tag{8a}$$

$$T = Ce^{-\kappa^2\alpha^2 t}. \tag{8b}$$

However, observe that (8a) is the general solution of (7a) only in the event that $\kappa \neq 0$, for if $\kappa = 0$ then the $\sin\kappa x$ term drops out. Since we do not yet know the value(s) of κ, we must allow for the possibility that the value $\kappa = 0$ will be needed. Setting $\kappa = 0$ in (7a) gives $X'' = 0$, with the general solution $X = D + Ex$, so we replace (8a) by the two-tier statement

$$X = \begin{cases} A\cos\kappa x + B\sin\kappa x, & \kappa \neq 0 \\ D + Ex, & \kappa = 0. \end{cases} \tag{9}$$

Apparently, we don't need to revise (8b) the way we revised (8a) because (8b) is a general solution of (7b) whether κ is zero or not. However, having already comitted ourselves in (9) to the separate treatment of these two cases, we replace (8b) by the two-tier statement †

$$T = \begin{cases} Fe^{-\kappa^2\alpha^2 t}, & \kappa \neq 0 \\ G, & \kappa = 0. \end{cases} \tag{10}$$

Thus far we have discussed the product solutions

$$u = XT = (D + Ex)G \tag{11}$$

* An alternative approach that you might prefer is to take $\partial/\partial x$ of (5), which step gives $F'(x) = 0$, so $F(x) =$ constant. From (5), $G(t)$ must also be constant, the same constant. [Or, $\partial/\partial t$ of (5) gives $G'(t) = 0$, so $G(t) =$ constant, and so on.] However, this approach is a bit weaker because it requires an assumption that X''/X is differentiable, an assumption that is not needed.

† $F\exp(-\kappa^2\alpha^2 t)$ can be expressed as $F\exp(-t/\tau)$, where $\tau = 1/(\kappa^2\alpha^2)$ is a **time constant**, namely, the time it takes the exponential to decay by 63% (from 1 to e^{-1}).

corresponding to $\kappa = 0$, and

$$u = XT = (A\cos\kappa x + B\sin\kappa x)Fe^{-\kappa^2\alpha^2 t} \tag{12}$$

for any $\kappa \neq 0$. Since D, E, G, A, B, F are arbitrary constants, we can combine DG as H and EG as I and simplify (11) as

$$u = H + Ix$$

for $\kappa = 0$, and we can combine AF as J and BF as K and simplify (12) as

$$u = (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2\alpha^2 t}$$

for $\kappa \neq 0$. Since (1a) is linear, the sum of these solutions must also be a solution, so we can write

$$u = H + Ix + (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2\alpha^2 t}, \tag{13}$$

where the constants H, I, J, K, and κ are arbitrary. But it is understood that $\kappa \neq 0$ in (13) because the $H+Ix$ part of (13) already accounts for the $\kappa = 0$ case. By the linearity of (1a) we can superimpose any number of such terms for different values of κ. With $\kappa = 1, 2$, and $\sqrt{5}$, for instance, we can write

$$\begin{aligned} u(x,t) &= (H + Ix) + (J_1\cos x + K_1\sin x)e^{-\alpha^2 t} \\ &\quad + (J_2\cos 2x + K_2\sin 2x)e^{-4\alpha^2 t} + (J_3\cos\sqrt{5}x + K_3\sin\sqrt{5}x)e^{-5\alpha^2 t}. \end{aligned} \tag{14}$$

The latter expression satisfies (1a) because each term does, and because (1a) is linear. [We urge you to verify that (14) satisfies (1a), by direct substitution.] But since we do not yet know what κ values to choose, let us continue with the more compact form (13).

We are ready to apply the boundary conditions (1b) and the initial condition (1c). Beginning with the left end condition, (13) gives

$$u(0,t) = u_1 = H + Je^{-\kappa^2\alpha^2 t} \qquad (0 < t < \infty)$$

or, in a more suggestive form,

$$(H - u_1)(1) + J(e^{-\kappa^2\alpha^2 t}) = 0. \qquad (0 < t < \infty) \tag{15}$$

Since the functions 1 and $\exp(-\kappa^2\alpha^2 t)$ are linearly independent on the t interval,* it follows from (15) that we need $H - u_1 = 0$ (so $H = u_1$) and $J = 0$. Updating (13) to incorporate these results we have, thus far,

$$u(x,t) = u_1 + Ix + K\sin\kappa x\, e^{-\kappa^2\alpha^2 t}. \tag{16}$$

Applying the right end condition next, (16) gives

$$u(L,t) = u_2 = u_1 + IL + K\sin\kappa L\, e^{-\kappa^2\alpha^2 t}$$

or

$$(IL + u_1 - u_2)(1) + K\sin\kappa L\, e^{-\kappa^2\alpha^2 t} = 0. \tag{17}$$

*Recall that *two* functions are linearly dependent if and only if one is a scalar multiple of the other, and neither 1 nor $\exp(-\kappa^2\alpha^2 t)$ is a scalar multiple of the other.

Again invoking the linear independence of 1 and $\exp(-\kappa^2\alpha^2 t)$, it follows from (17) that

$$IL + u_1 - u_2 = 0 \tag{18a}$$

and

$$K \sin \kappa L = 0. \tag{18b}$$

Equation (18a) gives $I = (u_2 - u_1)/L$, but (18b) presents a choice: either $K = 0$ or $\sin \kappa L = 0$ (or both). Here, and at analogous points in examples to follow, the rule is to *make that choice so as to maintain as robust a solution as possible* [because we still have the initial condition $u(x, 0) = f(x)$ to satisfy, and we will need all the help we can get]. If we choose $K = 0$, then we lose the $\sin \kappa x \exp(-\kappa^2\alpha^2 t)$ term in (16) and are left with

$$u(x,t) = u_1 + (u_2 - u_1)\frac{x}{L}, \tag{19}$$

which is capable of satisfying the initial condition only in the unlikely event that $f(x)$ happens to be $u_1 + (u_2 - u_1)(x/L)$. However, if we choose

$$\sin \kappa L = 0, \tag{20}$$

then K need not be zero, we retain the $\sin \kappa x \exp(-\kappa^2\alpha^2 t)$ term in (16), and (20) serves to identify the allowable values of κ, namely,

$$\kappa = \frac{n\pi}{L} \tag{21}$$

for $n = 0, \pm 1, \pm 2, \dots$.[*] Of these values we discard $n = 0$ because it gives $\kappa = 0$, whereas it was understood that the κ's in (16) were to be nonzero. Further, $n = -1, -2, \dots$ can be discarded since the $K \sin(n\pi x/L) \exp[-(n\pi\alpha/L)^2 t]$ combination in (16) is insensitive to the sign of n – to within a factor of ± 1, which factor can be absorbed by K anyhow.[†]

Using superposition as in (14) but for the values $\kappa = n\pi/L$ $(n = 1, 2, \dots)$, (16) gives

$$u(x,t) = u_1 + (u_2 - u_1)\frac{x}{L} + \sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}. \tag{22}$$

Before completing the solution let us review where we stand. The right-hand side of (22) satisfies the boundary conditions (1b) because $\sin(n\pi x/L) = 0$ at $x = 0$ and at $x = L$, for each $n = 1, 2, \dots$. Further, it appears to satisfy the PDE $L[u] = 0$ because it is a linear combination of product solutions

$$u_1 + (u_2 - u_1)\frac{x}{L}, \quad \sin\frac{\pi x}{L}\, e^{-(\pi\alpha/L)^2 t}, \quad \sin\frac{2\pi x}{L}\, e^{-(2\pi\alpha/L)^2 t}, \quad \dots,$$

each of which satisfies $L[u] = 0$. That is, if $L[\phi_1(x,t)] = 0, \dots, L[\phi_k(x,t)] = .0$, then $L[\alpha_1\phi_1(x,t) + \cdots + \alpha_k\phi_k(x,t)] = 0$ too because

$$L[\alpha_1\phi_1 + \cdots + \alpha_k\phi_k] = \alpha_1 L[\phi_1] + \cdots + \alpha_k L[\phi_k] \tag{23}$$

[*] Surely, $\sin z = 0$ has the roots $z = n\pi$ on the real axis. In fact, even if we broaden the search and look in the complex z plane, we find only the roots $n\pi$ on the real axis.

[†] Put differently, observe that only κ^2 appears in (7a) and (7b). Since positive and negative values of κ are therefore indistinguishable in the ODE's, their general solutions for a given positive κ, say $\kappa = \kappa_0$, must be the same as their general solutions for $\kappa = -\kappa_0$.

by virtue of the linearity of L. However, k in (23) is finite, whereas k in (22) is infinite, so the step

$$L\left[\sum_{n=1}^{\infty}\alpha_n\phi_n\right]=\sum_{n=1}^{\infty}\alpha_n L\left[\phi_n\right] \tag{24}$$

amounts to an interchange in the order of two limit operations, the derivative in L and the infinite series, and that step needs to be rigorously justified. This point of rigor comes up in Chapter 17 when we use Fourier series solution forms to satisfy ODE's (e.g., Comment 2 of Example 3 in Section 17.3), and is addressed in the optional Section 17.5. We address it here too, in the optional subsection 18.3.2. Generally, however, in Chapters 18–20, we omit rigorous justification and are satisfied with **formal solutions**.

Finally, we impose the initial condition (1c) on (22):

$$u(x,0)=f(x)=u_1+(u_2-u_1)\frac{x}{L}+\sum_{n=1}^{\infty}K_n\sin\frac{n\pi x}{L}. \qquad (0<x<L) \tag{25}$$

Two questions arise: given $f(x)$, is it *possible* to find coefficients K_n so as to satisfy (25) on $0<x<L$ and, if it is possible, then *how* do we determine the K_n's? First, put all the known terms on the left-hand side:

$$f(x)-u_1-(u_2-u_1)\frac{x}{L}=\sum_{n=1}^{\infty}K_n\sin\frac{n\pi x}{L}, \qquad (0<x<L) \tag{26}$$

and let us denote $f(x)-u_1-(u_2-u_1)(x/L)$ as $F(x)$, for brevity. Observing that

$$F(x)=\sum_{n=1}^{\infty}K_n\sin\frac{n\pi x}{L} \qquad (0<x<L) \tag{27}$$

is of the form of a half-range sine expansion of F, we can conclude that if F is sufficiently well-behaved,* then (27) is indeed possible, and [according to (4) in Section 17.4] the K_n's are computed as

$$K_n=\frac{2}{L}\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx. \tag{28}$$

Thus, our formal solution consists of (22), with the K_n's computed from (28).

To illustrate, consider a 10 cm-long copper rod held in boiling water until its temperature is 100° C throughout. At time $t=0$ it is removed and its ends are quenched with ice for all $t>0$. (We can either neglect heat loss from its lateral surface or specify that that surface be insulated.) Then $\alpha^2=1.14\,\text{cm}^2/\text{sec}$ (for copper), $L=10\,\text{cm}$, $u_1=u_2=0$, and $F(x)=100°$ C, so (28) gives

$$K_n=\frac{2}{10}\int_0^{10}100\sin\frac{n\pi x}{10}\,dx=\begin{cases}\dfrac{400}{n\pi}, & n \text{ odd}\\ 0, & n \text{ even}\end{cases} \tag{29}$$

*Let $F_{\text{ext}}(x)$ be the extended version of $F(x)$ that is $2L$-periodic and symmetric about $x=0$ and $x=L$. If $F_{\text{ext}}(x)$ and $F'_{\text{ext}}(x)$ are piecewise continuous on $[-L,L]$, then (27) converges in the sense of Theorem 17.3.1.

and (22) becomes

$$u(x,t) = \frac{400}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin \frac{n\pi x}{10}\, e^{-0.1125n^2 t}. \tag{30}$$

We have summed the series in (30) at a number of x's and at several representative times and have plotted the results in Fig. 2.

Since this is our first example there are numerous points to clarify, and we address them in the following comments.

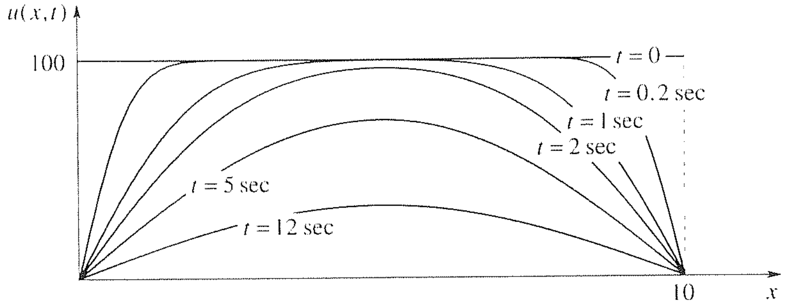


Figure 2. $u(x,t)$ for the case where $\alpha^2 = 1.14$, $L = 10$, $u_1 = u_2 = 0$, and $f(x) = 100$.

COMMENT 1. Notice the physical significance of the terms in (22). As $t \to \infty$ the exponential terms tend to zero, leaving the **steady-state** solution

$$\lim_{t\to\infty} u(x,t) \equiv u_s(x) = u_1 + (u_2 - u_1)\frac{x}{L}, \tag{31}$$

which is a linear variation from u_1 at $x = 0$ to u_2 at $x = L$ (Fig. 3). Thus, the summation term in (22) represents the **transient** part of the solution, that links the initial distribution $u(x,0) = f(x)$ to the steady-state distribution $u_s(x)$. For the specific case represented in Fig. 2, the steady-state solution is simply $u_s(x) = 0$ because $u_1 = u_2 = 0$.

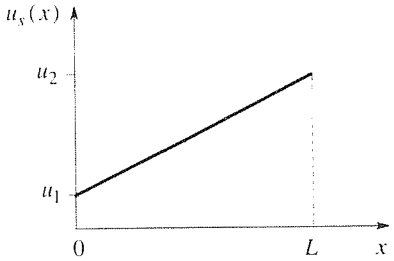


Figure 3. Steady-state.

COMMENT 2. Notice further that within the transient part of (22) each term dies out exponentially faster (as $t \to \infty$) than the preceding term because of the n^2 in the exponent. To illustrate the significance of that point let us write out (30) at the representative times $t = 0.2$, 1, and 12 seconds:

$$\begin{aligned}
u(x,0.2) &= 124.5 \sin\frac{\pi x}{10} + 34.7 \sin\frac{3\pi x}{10} + 14.5 \sin\frac{5\pi x}{10} + \cdots,\\
u(x,1) &= 113.8 \sin\frac{\pi x}{10} + 15.4 \sin\frac{3\pi x}{10} + 1.5 \sin\frac{5\pi x}{10} + \cdots,\\
u(x,12) &= 33.0 \sin\frac{\pi x}{10} + 0.0002 \sin\frac{3\pi x}{10} + 6\times 10^{-14} \sin\frac{5\pi x}{10} + \cdots,
\end{aligned}$$

which results are among those shown in Fig. 2. Computationally speaking, then, (30) [more generally, (22)] is an excellent result for "large" times because if t is large enough then only one or two of the terms in the series are needed for engineering accuracy. Conversely, for "small" times a great many terms in the series may be needed. For instance, at $t = 0$ the terms in (30) die out only like $1/n$ because the exponential factor is unity at $t = 0$.

COMMENT 3. Why did we write $-\kappa^2$ in (16), rather than $+\kappa^2$? That is, how did we anticipate that the separation constant would be negative? The rule of thumb is to choose either sign and then examine the resulting ODE's for clues. Specifically, the minus sign chosen in (6) causes plus signs in (7a) and (7b), both of which look good: the plus sign in (7a) looks good because it results in sine and cosine solutions, which will be needed for the eventual Fourier series expansion of $f(x) - u_s(x)$, and the plus sign in (7b) looks good because it results in exponential decay rather than physically unreasonable exponential growth. (See Exercise 2.)

COMMENT 4. It is essential to *apply the boundary conditions before the initial condition*. To see why, let us try imposing the initial condition first, instead. Then, (13) gives

$$u(x,0) = f(x) = H + Ix + J\cos\kappa x + K\sin\kappa x,$$

which cannot be satisfied unless $f(x)$ happens to be a linear combination of 1, x, $\cos\kappa x$, and $\sin\kappa x$, for some κ. Applying the boundary conditions first enabled us to obtain the solution form (22), which form was powerful enough to handle any given initial condition $u(x,0) = f(x)$. This sequencing – boundary conditions first and initial condition second – will be appropriate all through Chapters 18 and 19.

COMMENT 5. As emphasized in Section 18.2.1, we did *not* find a *general solution* of the diffusion equation and then apply the boundary and initial conditions. Rather, we used the method of separation of variables to develop a solution that was sufficiently robust to handle the boundary conditions and initial condition.

COMMENT 6. It is interesting that the assumed product form $u(x,t) = X(x)T(t)$ seems a bad choice because it maintains the same shape, $X(x)$, for all time and is merely scaled in magnitude by the time-varying factor $T(t)$. Rather, we would expect the shape of $u(x,t)$ to change with time – as in Fig. 2, for instance, where $u(x,t)$ is initially a constant but approaches a sinusoidal shape as heat diffuses out of the rod at both ends due to the end conditions $u(0,t) = u(L,t) = 0$. However, understand that the *superposition* of product solutions is not itself a product solution; that is, the final solution (22) is *not* a function of x times a function of t.

COMMENT 7. Still retaining the boundary conditions $u(0,t) = u(L,t) = 0$, what would happen if we changed the initial condition from $u(x,0) = 100$ to

$$u(x,0) = \begin{cases} 100, & 0 < x < x_0 \text{ and } x_0 < x < L \\ 5\times 10^6, & x = x_0 \end{cases} \tag{32}$$

that is, if we changed the initial temperature at a single point x_0 to $5,000,000°$? Nothing; the solution would still be given by (30), because a change in the integrand of (28) at a single point, from one finite value to another, cannot change the value of the integral.

COMMENT 8. The solution of (1) is not necessarily an infinite series. For instance, suppose that $u_1 = u_2 = 0$ and

$$u(x,0) = f(x) = 40\sin\frac{\pi x}{L}. \tag{33}$$

Then, application of the initial condition (33) to (22) gives

$$\begin{aligned}40\sin\frac{\pi x}{L} &= \sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L}\\ &= K_1\sin\frac{\pi x}{L} + K_2\sin\frac{2\pi x}{L} + \cdots,\end{aligned}$$

which is satisfied, as can be seen by inspection, by setting $K_1 = 40$ and all the other K_n's equal to zero. Thus, we obtain the one-term solution

$$u(x,t) = 40\sin\frac{\pi x}{L}e^{-(\pi\alpha/L)^2 t}. \tag{34}$$

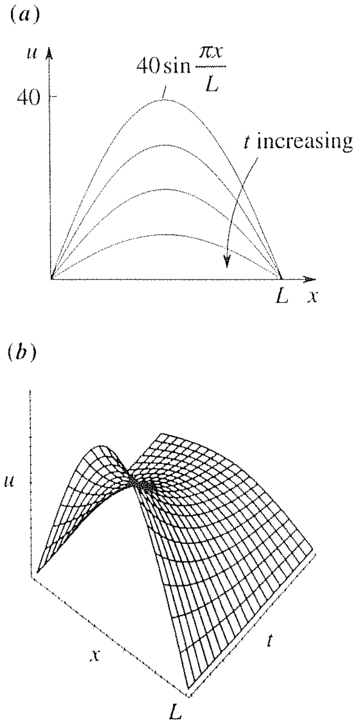


Figure 4. Graph of (34).

We can display (34) in two dimensions by plotting u versus x at representative times, as we have done in Fig. 4a [and as we did earlier in Fig. 2 for the case where $f(x)$ was 100], or in three dimensions by plotting the u surface above the x, t plane, as we have done in Fig. 4b. Similarly, if

$$u(x,0) = f(x) = 30\sin\frac{2\pi x}{L} - 25\sin\frac{5\pi x}{L}, \tag{35}$$

then

$$u(x,t) = 30\sin\frac{2\pi x}{L}e^{-(2\pi\alpha/L)^2 t} - 25\sin\frac{5\pi x}{L}e^{-(5\pi\alpha/L)^2 t}, \tag{36}$$

and so on. ▮

In dwelling on so many details, in Example 1, we have shown no mercy. However, Example 1 provides the basic model for the application of the method of separation of variables, as it is used here in Chapters 18–20, so careful study of that example is well worth the effort.

EXAMPLE 2. *Different Boundary Conditions.* This time suppose that the left end of the rod is insulated, and the right end is held at a constant temperature 100 for all $t > 0$, so the problem is as follows:

$$L[u] = \alpha^2 u_{xx} - u_t = 0, \qquad (0 < x < L,\ 0 < t < \infty) \tag{37a}$$

$$u_x(0,t) = 0, \quad u(L,t) = 100, \qquad (0 < t < \infty) \tag{37b}$$

$$u(x,0) = f(x), \qquad (0 < x < L) \tag{37c}$$

and suppose f is the piecewise constant function

$$f(x) = \begin{cases} 60, & 0 < x < L/2\\ 0, & L/2 < x < L\end{cases} \tag{38}$$

shown in Fig. 5.

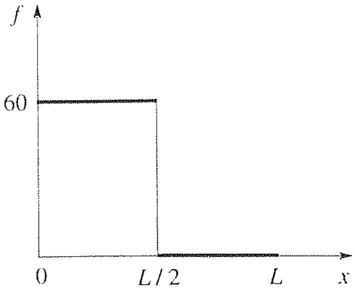


Figure 5. $f(x)$.

Up until (13) the story is unchanged, so let us begin with

$$u(x,t) = H + Ix + (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2\alpha^2 t}. \tag{39}$$

Applying the left-hand boundary condition,

$$u_x(x,t) = I + (-\kappa J \sin \kappa x + \kappa K \cos \kappa x)e^{-\kappa^2 \alpha^2 t},$$

so

$$u_x(0,t) = 0 = I + \kappa K e^{-\kappa^2 \alpha^2 t}, \tag{40}$$

which gives $I = K = 0$. Updating (39) accordingly,

$$u(x,t) = H + J \cos \kappa x \, e^{-\kappa^2 \alpha^2 t}. \tag{41}$$

Next, the right-hand boundary condition gives

$$u(L,t) = 100 = H + J \cos \kappa L \, e^{-\kappa^2 \alpha^2 t}, \tag{42}$$

so $H = 100$ and $J \cos \kappa L = 0$. We cannot afford to satisfy the latter by setting $J = 0$ because if $J = 0$ then we lose the $\cos \kappa L e^{-\kappa^2 \alpha^2 t}$ term in (41), and the latter reduces to $u(x,t) = H = 100$, which cannot satisfy the initial condition. Rather, we satisfy $J \cos \kappa L = 0$ by setting

$$\cos \kappa L = 0 \tag{43}$$

and letting J remain arbitrary. Thus, $\kappa L = \pi/2, 3\pi/2, 5\pi/2, \ldots$, so

$$\kappa = \frac{n\pi}{2L}, \qquad (n = 1, 3, \ldots) \tag{44}$$

and we have

$$u(x,t) = 100 + \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L} e^{-(n\pi\alpha/2L)^2 t}. \tag{45}$$

Finally, the initial condition is

$$u(x,0) = f(x) = 100 + \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L}$$

or

$$f(x) - 100 = \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L}. \qquad (0 < x < L) \tag{46}$$

Comparing the form of the right-hand side of (46) with the half- and quarter-range cosine and sine expansion formulas, we see that (46) amounts to a quarter-range cosine expansion of $f(x) - 100$, so

$\hookrightarrow$ pág 871

$$\begin{aligned} J_n &= \frac{2}{L} \int_0^L [f(x) - 100] \cos \frac{n\pi x}{2L} \, dx \\ &= \frac{2}{L} \left[\int_0^{L/2} (-40) \cos \frac{n\pi x}{2L} \, dx + \int_{L/2}^L (-100) \cos \frac{n\pi x}{2L} \, dx \right] \\ &= \frac{80}{n\pi} \left(3 \sin \frac{n\pi}{4} - 5 \sin \frac{n\pi}{2} \right). \end{aligned} \tag{47}$$

The solution is given by (45) and (47).

To plot the results, let the material be glass (so $\alpha^2 = 0.0034\,\text{cm}^2/\text{sec}$), and let $L = 10\,\text{cm}$. In Fig. 6 we plot u versus x at representative times.

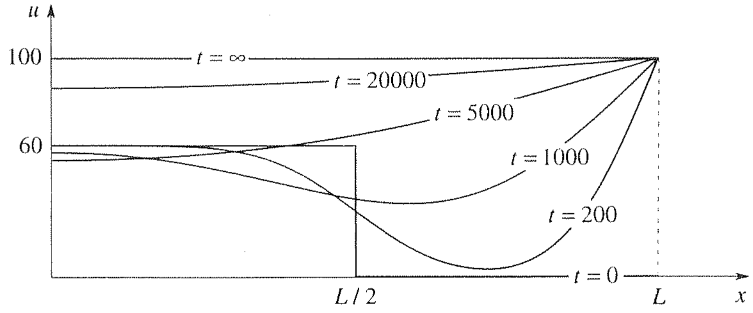


Figure 6. Graph of (45) for $\alpha^2 = 0.0034$ and $L = 10$.

COMMENT 1. Notice that each graph of u is flat at $x = 0$, in accord with the boundary condition $u_x(0,t) = 0$, and that the steady-state solution $u_s(x) = 100$ is approached as $t \to \infty$. That the initial temperature $u(x,0) = f(x)$ also satisfies the boundary condition $u_x(0,t) = 0$ is only by coincidence and is not required since we require the boundary conditions to hold only for $0 < t < \infty$, not for $0 \le t < \infty$.

COMMENT 2. You might be tempted to break (46) into two parts,

$$-40 = \sum_{n=1,3,\dots}^{\infty} J_n \cos\frac{n\pi x}{2L} \qquad \left(0 < x < \frac{L}{2}\right) \tag{48a}$$

and

$$-100 = \sum_{n=1,3,\dots}^{\infty} J_n \cos\frac{n\pi x}{2L}, \qquad \left(\frac{L}{2} < x < L\right) \tag{48b}$$

and to use the quarter-range cosine expansion formulas to solve for the J_n's in each case. That procedure would be INCORRECT for two reasons. First, it will give J_n's that have different values in the intervals $0 < x < L/2$ and $L/2 < x < L$, in which case they will be functions of x, not constants – as they were supposed to be. Second, we cannot use the quarter-range cosine formulas to solve (48a) and (48b) for the J_n's because $\sum_{n=1,3,\dots}^{\infty} J_n \cos(n\pi x/2L)$ is a quarter-range cosine expansion on $0 < x < L$, not on $0 < x < L/2$ or on $L/2 < x < L$. ▮

In Example 1, $u(0,t)$ and $u(L,t)$ are prescribed constants and we end up expanding $f(x) - u_s(x)$ in a half-range sine series. In Example 2, $u_x(0,t)$ and $u(L,t)$ are prescribed constants and we use a quarter-range cosine series. You will find that if instead $u(0,t)$ and $u_x(L,t)$ are prescribed constants, then a quarter-range sine series will be appropriate. If $u_x(0,t)$ and $u_x(L,t)$ are prescribed constants that are equal, then a half-range cosine series will be appropriate, but if they are unequal then the solution is more difficult (Exercise 19).

18.3.2. Verification of solution. (Optional) As stated, we can claim only to have found *formal* solutions of the diffusion problems in Examples 1 and 2. To illustrate the verification process let us verify that the formal solution (22) of the problem (1), with the K_n's given by (28), does satisfy the requirements in (1).

First let us show that (22) satisfies the PDE $\alpha^2 u_{xx} = u_t$. If we differentiate (22) termwise we obtain

$$u_x = \frac{u_2 - u_1}{L} + \sum_{n=1}^{\infty} \frac{n\pi}{L} K_n \cos\frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}, \tag{49a}$$

$$u_{xx} = -\sum_{n=1}^{\infty} \left(\frac{n\pi}{L}\right)^2 K_n \sin\frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}, \tag{49b}$$

$$u_t = -\sum_{n=1}^{\infty} \left(\frac{n\pi\alpha}{L}\right)^2 K_n \sin\frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}, \tag{49c}$$

so we see that $\alpha^2 u_{xx}$ does equal u_t, provided that we can rigorously justify the termwise differentiations that produced (49a,b,c). Theorem 17.5.2 tells us that those steps are permissible if the resulting series [on the right-hand sides of (49a,b,c)] converge *uniformly* on the problem domain $0 < x < L$, $0 < t < \infty$, and Theorem 17.5.1 gives us the Weierstrass M-test as a test for uniform convergence. To apply the latter to the series in (49a), note that

$$\left|\frac{n\pi}{L} K_n \cos\frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}\right| \le Qne^{-(n\pi\alpha/L)^2 t_0} \equiv M_n \tag{50}$$

for all $t > t_0$ and $0 < x < L$ because $|\cos(n\pi x/L)| \le 1$ and $K_n \to 0$ as $n \to \infty$,* so that the K_n's must be bounded: thus there must exist a finite positive constant Q such that $|\pi K_n/L| \le Q$ for all n's. It follows easily from the ratio test that $\sum_{n=1}^{\infty} M_n$ converges, so the series in (49a) does indeed converge uniformly in $0 < x < L$, $t_0 \le t < \infty$, for arbitrarily small t_0. Similarly for the series in (49b) and (49c), the only difference being that those series contain a factor of n^2 rather than n; but the ratio test shows that $\sum_{n=1}^{\infty} n^2 \exp\left[-(n\pi\alpha/L)^2 t_0\right]$ converges, just as $\sum_{n=1}^{\infty} n \exp\left[-(n\pi\alpha/L)^2 t_0\right]$ does.

Turning to the boundary conditions (1b) and the initial condition (1c), it might appear that the satisfaction of these conditions does not even need verification. For instance, is not

$$u(0,t) = u_1 + 0 + \sum_{n=1}^{\infty} 0 = u_1$$

obviously true? The point to emphasize is that in posing PDE problems we require satisfaction of the PDE in the *open* region, in this case $0 < x < L$, $0 < t < \infty$. We do that so that we can allow for not-so-well-behaved boundary and initial data. For if $\alpha^2 u_{xx} = u_t$ were to be satisfied in the region $0 \le x \le L$, $0 \le t < \infty$, then we would need $u(x,0) = f(x)$ to be twice differentiable, and $u(0,t) = g(t)$

*See the last sentence of Exercise 1, Section 17.6.

and $u(L,t) = h(t)$, say, to be once differentiable, whereas we want to allow for boundary and initial data that are not even continuous, let alone differentiable. In Example 2, for instance, $f(x)$ is discontinuous. The way we link the solution $u(x,t)$ in the open domain to the boundary and initial conditions is through limits. That is, by (1b) we mean

$$\lim_{x\to 0+} u(x,t) = u_1 \quad \text{and} \quad \lim_{x\to L-} u(x,t) = u_2, \qquad (0 < t < \infty) \tag{51}$$

and by (1c) we mean

$$\lim_{t\to 0+} u(x,t) = f(x). \qquad (0 < x < L) \tag{52}$$

To verify that (22) satisfies (51) we can use the following theorem.

THEOREM 18.3.1 *Continuity of Sum Function*
If $\sum_{n=1}^{\infty} a_n(x)$ converges uniformly to $s(x)$ on some x interval I and the $a_n(x)$'s are continuous on I, then $s(x)$ is continuous on I.

That is, if the convergence is uniform, then the continuity of the partial sums is passed on to the sum function $s(x)$.

Applying that result, observe first that the series in (22) converges uniformly on $0 \leq x \leq L$ for each t such that $0 < t_0 \leq t < \infty$ because there is a finite constant P such that

$$\left| K_n \sin \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t} \right| \leq P e^{-(n\pi\alpha/L)^2 t_0} \equiv M_n$$

there, and $\sum_{n=1}^{\infty} M_n$ is convergent. Thus the right-hand side of (22) is a continuous function of x on $0 \leq x \leq L$, so its values at $x = 0$ and at $x = L$, namely u_1 and u_2, respectively, are the same as their limiting values as $x \to 0+$ and as $x \to L-$, respectively.

Verification of the initial condition (52) can be accomplished with the help of a theorem of Abel. For that, and a generally more detailed discussion, we refer you to Churchill and Brown.*

Finally, there is the question of uniqueness: is (22) the *only* solution of (1)? A formal proof of uniqueness is outlined in Exercise 25; for detailed discussion we refer you, again, to Churchill and Brown.

18.3.3. Use of Sturm–Liouville theory. (Optional) In subsection 18.3.1 we found that the final step of the separation-of-variables solution involves the expansion of a given function. For the diffusion equation $\alpha^2 u_{xx} = u_t$ with Dirichlet or Neumann boundary conditions that expansion was a half- or quarter-range cosine or

*R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed. (New York: McGraw Hill, 1978, p. 129).

sine series, so we were guided by our knowledge of such series. More generally, the necessary expansion is an *eigenfunction expansion* in terms of the orthogonal eigenfunctions of a Sturm–Liouville problem that is "built in." Thus, a more powerful approach is to appeal to the Sturm–Liouville theory, which includes the half- and quarter-range expansions as special cases. Let us illustrate with three examples.

EXAMPLE 3. *Example 1 Reconsidered.* The Sturm–Liouville theory becomes relevant only when we reach the point of needing to carry out the expansion of a given function. Thus, in reconsidering Example 1 in the light of the Sturm–Liouville theory we can begin with equation (26). Our claim is that the $\sin(n\pi x/L)$ functions in (26) are the orthogonal eigenfunctions of a Sturm–Liouville problem governing $X(x)$, namely, the equation (7a) together with the homogeneous boundary conditions $X(0) = 0$ and $X(L) = 0$:

$$X'' + \kappa^2 X = 0, \qquad (0 < x < L) \tag{53a}$$

$$X(0) = 0, \quad X(L) = 0, \tag{53b}$$

which problem is indeed of the Sturm–Liouville form

$$(py')' + qy + \lambda wy = 0, \qquad (a < x < b) \tag{54a}$$

$$\alpha y(a) + \beta y'(a) = 0, \quad \gamma y(b) + \delta y'(b) = 0, \tag{54b}$$

where $y(x)$ is $X(x)$, $p(x) = w(x) = 1$, $q(x) = 0$, $\lambda = \kappa^2$, $a = 0$, $b = L$, $\alpha = \gamma = 1$, and $\beta = \delta = 0$. Considering the boundary conditions $u(0,t) = u_1$ and $u(L,t) = u_2$, where did we get $X(0) = 0$ and $X(L) = 0$, in (53b)? Recall that the $u_1 + (u_2 - u_1)x/L$ terms in (22) make up the steady-state solution $u_s(x)$. The burden of satisfying the nonhomogeneous boundary conditions, $u(0,t) = u_1$ and $u(L,t) = u_2$, is carried by the steady-state solution $u_s(x)$, so the transient part, $\sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L} \exp[-(n\pi\alpha/L)^2 t]$, must be *zero* at $x = 0$ and $x = L$. Thus, the $X(x) = \sin(n\pi x/L)$ eigenfunctions, contained therein, actually satisfy the *homogeneous conditions* (53b), as is easily verified by evaluating $\sin(n\pi x/L)$ at $x = 0$ and $x = L$.

According to the Sturm–Liouville theory, then, (27) can indeed be satisfied, and the expansion coefficients are given by

$$K_n = \frac{\langle F(x), \sin\frac{n\pi x}{L}\rangle}{\langle \sin\frac{n\pi x}{L}, \sin\frac{n\pi x}{L}\rangle} = \frac{\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx}{\int_0^L \sin^2\frac{n\pi x}{L}\,dx}$$

$$= \frac{2}{L}\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx, \tag{55}$$

which result is the same as (28). [Recall that the weight function, in the two integrals, in this case is $w(x) = 1$.] ▮

Thus, the idea is that satisfaction of the initial condition $u(x,0) = f(x)$ requires the expansion of $F(x) = f(x) - u_s(x)$. The latter will inevitably be an eigenfunction expansion in terms of the eigenfunctions $\phi_n(x)$ [namely, $\sin(n\pi x/L)$

in Example 3] of a Sturm–Liouville problem on $X(x)$. The Sturm–Liouville theory assures us that the desired expansion is possible, and it tells us how to compute the expansion coefficients – namely, as $\langle F(x), \phi_n\rangle / \langle \phi_n, \phi_n\rangle$, where the inner product has weight function $w(x)$.

In Example 3 we were able to use either the half-range sine expansion concept *or* the Sturm–Liouville theory. In the next example the expansion will not be of half- or quarter-range type, so we will *have to* use the Sturm–Liouville theory.

EXAMPLE 4. This time consider the problem

$$L[u] = u_{xx} - u_t = 0, \qquad (0 < x < 1,\ 0 < t < \infty) \tag{56a}$$

$$u(0,t) - 2u_x(0,t) = 5, \quad u(1,t) = 35, \qquad (0 < t < \infty) \tag{56b}$$

$$u(x,0) = f(x), \qquad (0 < x < 1) \tag{56c}$$

where we have set $\alpha^2 = 1$ and $L = 1$ for simplicity. Observe that $u(0,t) - 2u_x(0,t) = 5$ is a Robin boundary condition, a boundary condition of the third kind.

Separating variables as usual, let us begin with equation (13):

$$u(x,t) = H + Ix + (J \cos \kappa x + K \sin \kappa x)e^{-\kappa^2 t}. \tag{57}$$

Applying the left end condition gives

$$\begin{aligned} u(0,t) - 2u_x(0,t) = 5 &= H + Je^{-\kappa^2 t} - 2(I + \kappa K e^{-\kappa^2 t}) \\ &= (H - 2I) + (J - 2\kappa K)e^{-\kappa^2 t}, \qquad (0 < t < \infty) \end{aligned}$$

so

$$H - 2I = 5, \tag{58a}$$

$$J - 2\kappa K = 0. \tag{58b}$$

And the right end condition gives

$$u(1,t) = 35 = H + I + (J \cos \kappa + K \sin \kappa)e^{-\kappa^2 t}, \qquad (0 < t < \infty)$$

so

$$H + I = 35, \tag{59a}$$

$$J \cos \kappa + K \sin \kappa = 0. \tag{59b}$$

Equations (58a) and (59a) give $H = 25$ and $I = 10$. Equations (58b) and (59b) give the unique trivial solution $J = K = 0$, unless we choose κ so that the determinant of the coefficient matrix vanishes:

$$\begin{vmatrix} 1 & -2\kappa \\ \cos \kappa & \sin \kappa \end{vmatrix} = 0. \tag{60}$$

We cannot accept the trivial solution $J = K = 0$ because it reduces (57) to $u(x,t) = 25 + 10x$, which does satisfy the PDE and boundary conditions but which cannot satisfy the initial condition $u(x,0) = f(x)$ unless $f(x)$ happens to be $25 + 10x$. Thus, we impose the determinant condition (60), or

$$\tan \kappa = -2\kappa, \tag{61}$$

and designate the positive roots of (61) as $\kappa_1, \kappa_2, \ldots$. Next, we need to *find* the corresponding nontrivial solutions of (58b) and (59b) for J and K. With (61) satisfied, (58b) and (59b) are redundant, so we can drop (59b) and use (58b) to obtain $J = 2\kappa K$.

With these results, and superposition, (57) gives

$$u(x,t) = 25 + 10x + \sum_{n=1}^{\infty} K_n \phi_n(x) e^{-\kappa_n^2 t}, \tag{62}$$

where

$$\phi_n(x) = 2\kappa_n \cos \kappa_n x + \sin \kappa_n x. \tag{63}$$

Finally, the initial condition requires that

$$u(x,0) = f(x) = 25 + 10x + \sum_{n=1}^{\infty} K_n \phi_n(x),$$

or,

$$F(x) = \sum_{n=1}^{\infty} K_n \phi_n(x), \qquad (0 < x < 1) \tag{64}$$

where $F(x) = f(x) - (25 + 10x)$, $25 + 10x$ being the steady-state solution $u_s(x)$. The ϕ_n's in (64) are the eigenfunctions generated by the Sturm–Liouville problem

$$X'' + \kappa^2 X = 0, \qquad (0 < x < 1) \tag{65a}$$

$$X(0) - 2X'(0) = 0, \quad X(1) = 0, \tag{65b}$$

with weight function $w(x) = 1$, so

$$K_n = \frac{\langle F, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} = \frac{\displaystyle\int_0^1 F(x)(2\kappa_n \cos \kappa_n x + \sin \kappa_n x)\, dx}{\displaystyle\int_0^1 (2\kappa_n \cos \kappa_n x + \sin \kappa_n x)^2\, dx}. \tag{66}$$

In fact, this Sturm–Liouville problem, including the determination of the κ_n's, is the subject of Example 3 in Section 17.7. ▮

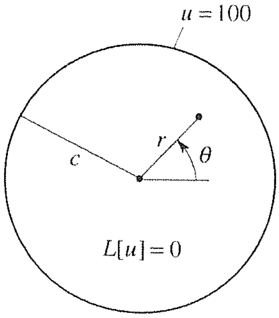


Figure 7. Conduction in a disk.

EXAMPLE 5. *Unsteady Conduction in a Disk.* Consider the unsteady conduction of heat in a circular disk such as a coin of radius c, the flat faces of which are insulated (Fig. 7). Then the temperature field $u(r,t)$ in the disk is governed by the problem

$$\alpha^2 \nabla^2 u = \alpha^2 \left(u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta} \right) = u_t, \qquad (0 \le r < c,\ 0 < t < \infty)$$
$$u(c,t) = 100, \quad u(r,0) = f(r). \tag{67}$$

That is, the disk is initially at a prescribed temperature $f(r)$, and then we hold the outer edge (with boiling water, for example) at $u = 100°$ for all $t > 0$. Because the domain is circular, and the initial and boundary temperature are independent of θ, it appears that the resulting temperature field u will be independent of θ as well, and will be a function only of r and t. Hence, we can strike out the $u_{\theta\theta}$ term, and the PDE reduces to

$$\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t. \tag{68}$$

To solve by separation of variables, seek u in the product form

$$u(r,t) = R(r)T(t). \tag{69}$$

Putting (69) into the PDE and dividing by RT (and α^2) gives

$$\frac{R'' + \frac{1}{r}R'}{R} = \frac{1}{\alpha^2}\frac{T'}{T} = \text{constant} = -\kappa^2, \tag{70}$$

and hence the ODE's

$$R'' + \frac{1}{r}R' + \kappa^2 R = 0, \tag{71a}$$

$$T' + \kappa^2\alpha^2 T = 0. \tag{71b}$$

Equation (71a) is the subject of Example 1 in Section 4.6, and its general solution is

$$R(r) = AJ_0(\kappa r) + BY_0(\kappa r), \tag{72}$$

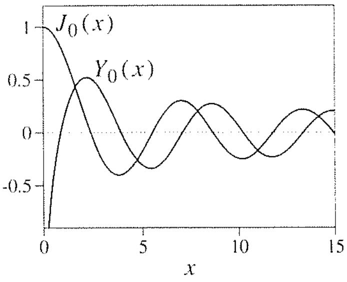


Figure 8. $J_0(x)$ and $Y_0(x)$.

where J_0, Y_0 are the Bessel functions of first and second kind, respectively, of order zero (Fig. 8). Are there any values of κ for which (71) fails to provide a general solution? Yes, for $\kappa = 0$ because $Y_0(0) = -\infty$ does not exist. But for $\kappa = 0$ (71a) can be expressed as $(rR')' = 0$, which can be integrated to give $R(r) = C + D\ln r$. Thus, we distinguish the cases $\kappa \neq 0$ and $\kappa = 0$, and write

$$R(r) = \begin{cases} AJ_0(\kappa r) + BY_0(\kappa r), & \kappa \neq 0 \\ C + D\ln r, & \kappa = 0 \end{cases} \tag{73}$$

$$T(t) = \begin{cases} Ee^{-\kappa^2\alpha^2 t}, & \kappa \neq 0 \\ F, & \kappa = 0. \end{cases} \tag{74}$$

Thus far, we have

$$\begin{aligned} u(r,t) &= (C + D\ln r)F + [AJ_0(\kappa r) + BY_0(\kappa r)]Ee^{-\kappa^2\alpha^2 t} \\ &= G + H\ln r + [PJ_0(\kappa r) + QY_0(\kappa r)]e^{-\kappa^2\alpha^2 t}, \end{aligned} \tag{75}$$

where we have combined CF as G, DF as H, and so on.

In the preceding examples, $X(x)$ is governed by a second-order ODE and there are two x boundary conditions (at $x = 0$ and at $x = L$). In the present example, likewise, $R(r)$ is governed by a second-order ODE, but we find only one r boundary condition in (67),

$u = 100$ at $r = c$. As a second r boundary condition it seems appropriate to require that u be bounded at $r = 0$.

As in preceding examples, we save the initial condition for last and begin by applying the boundary conditions. Of the two boundary conditions (u bounded at $r = 0$ and $u = 100$ at $r = c$), we recommend applying any boundedness condition first because it gives an immediate simplification. Specifically, for $u(0, t)$ to be bounded we need $H = Q = 0$ in (74) because both $\ln r$ and $Y_0(\kappa r)$ are unbounded at $r = 0$. Then (74) simplifies to

$$u(r, t) = G + PJ_0(\kappa r)e^{-\kappa^2\alpha^2 t}. \tag{76}$$

Next,

$$u(c, t) = 100 = G + PJ_0(\kappa c)e^{-\kappa^2\alpha^2 t},$$

or,

$$(G - 100)(1) + PJ_0(\kappa c)e^{-\kappa^2\alpha^2 t} = 0. \qquad (0 < t < \infty) \tag{77}$$

Since 1 and $\exp(-\kappa^2\alpha^2 t)$ are linearly independent on the t interval, it follows from (77) that $G - 100 = 0$ and $PJ_0(\kappa c) = 0$. The former gives $G = 100$ and the latter gives $P = 0$ or $J_0(\kappa c) = 0$. We reject the choice $P = 0$ because we cannot afford to lose the $PJ_0(\kappa r)\exp(-\kappa^2\alpha^2 t)$ term in (76), and adopt the choice

$$J_0(\kappa c) = 0 \tag{78}$$

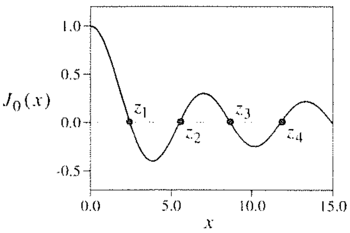


Figure 9. The zeros z_n of $J_0(x)$.

with positive roots $\kappa_n c = z_n$ for $n = 1, 2, \dots$, where the z_n's are the (known) zeros of J_0 (Fig. 9). With these choices, and the help of superposition, we have

$$u(r, t) = 100 + \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right) e^{-(z_n\alpha/c)^2 t}. \tag{79}$$

(The P_n's are arbitrary constants, not Legendre polynomials.)

Finally, the initial condition requires that

$$u(r, 0) = f(r) = 100 + \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right),$$

or

$$F(r) \equiv f(r) - 100 = \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right). \qquad (0 \le r < c) \tag{80}$$

The expansion functions $J_0(z_n r/c)$ are the eigenfunctions of the singular Sturm–Liouville problem

$$\begin{aligned} &(rR')' + \kappa^2 rR = 0, \qquad (0 < r < c) \\ &R(0) \text{ bounded}, \quad R(c) = 0, \end{aligned} \tag{81}$$

which problem is the subject of Example 2 in Section 17.8. Observe that we multiplied (71a) through by r in order to obtain the standard form given in (81). That step is important because it is only when the equation is in the standard form that we can identify from it the weight function – that will be needed in our inner product. From (81) we see that the weight function is $w(r) = r$, so we can write

$$P_n = \frac{\langle F(r), J_0(z_n r/c)\rangle}{\langle J_0(z_n r/c), J_0(z_n r/c)\rangle} = \frac{\displaystyle\int_0^c F(r) J_0\left(z_n \frac{r}{c}\right) r\, dr}{\displaystyle\int_0^c \left[J_0\left(z_n \frac{r}{c}\right)\right]^2 r\, dr}$$
$$= \frac{2}{c^2[J_1(z_n)]^2}\int_0^c F(r) J_0\left(z_n \frac{r}{c}\right) r\, dr. \tag{82}$$

For explanation of the last step, see Example 2 of Section 17.8.

COMMENT 1. As a concrete example, let $f(r) = 0$. Then (Exercise 31)

$$P_n = -\frac{200}{c^2[J_1(z_n)]^2}\int_0^c J_0\left(z_n \frac{r}{c}\right) r\, dr = -\frac{200}{z_n J_1(z_n)}, \tag{83}$$

so

$$u(r,t) = 100 - 200\sum_{n=1}^{\infty}\frac{1}{z_n J_1(z_n)} J_0\left(z_n \frac{r}{c}\right) e^{-(z_n\alpha/c)^2 t}. \tag{84}$$

For instance, the temperature history at the center of the disk is

$$u(0,t) = 100 - 200\sum_{n=1}^{\infty}\frac{1}{z_n J_1(z_n)} e^{-(z_n\alpha/c)^2 t}, \tag{85}$$

which is plotted in Fig. 10 for the case where the material is glass ($\alpha^2 = 0.0034\,\text{cm}^2/\text{sec}$ and $c = 10\,\text{cm}$).

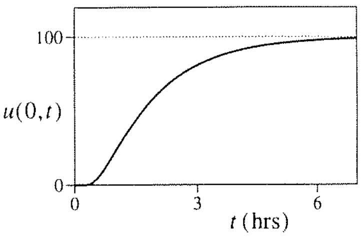


Figure 10. Temperature history at center; $\alpha^2 = 0.0034$, $c = 10$.

COMMENT 2. Observe that the point $r = 0$ is the left endpoint of the r interval $0 \le r < c$, but, physically, it corresponds to an interior point of the disk, the center of the disk. Thus, the PDE must be satisfied there. In particular, if $u_r(0,t)$ is to exist, then it must be zero. That is, the u surface (i.e., the graph of u above the r, θ plane) must be flat at $r = 0$ because otherwise the u surface will be conical there and $u_r(0,t)$ will not exist. Thus, in place of the condition that $R(0)$ be bounded we could have used the condition

$$R'(0) = 0, \tag{86}$$

and the latter would have produced the same results.

COMMENT 3. Recall that we argued that u does not vary with θ in this example. Hence, we dropped the $u_{\theta\theta}$ term in the PDE and solved the reduced equation $\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t$. If you still have doubts about that step, observe that (79) does satisfy the boundary conditions, the initial condition [if the P_n's are computed using (81)], and the *full* equation $\alpha^2\nabla^2 u = u_t$ including the $u_{\theta\theta}$ term because $\partial^2/\partial\theta^2$ of the right-hand side of (79) is zero. ▮

Closure. This section covers almost all aspects of the method of separation of variables, which method is used in each of Chapters 18–20. The starting point is to assume a procedure form for the solution, for instance $u(x,t) = X(x)T(t)$ if the independent variables are x and t. For the diffusion equation, that step enable us to separate the variables and obtain ODE's on $X(x)$ and $T(t)$. That simplification, from a PDE to two ODE's, is the point of the method.

The essential ingredients, for the success of the method in a given application, are that the equation needs to be separable, the boundary and initial conditions need to be given on constant coordinate curves, and the PDE needs to be linear – so that a sufficiently robust solution can be built up by the superposition of various product solutions. Only two PDE's are studied in the foregoing examples, $\alpha^2 u_{xx} = u_t$ and $\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t$, and these three conditions are met: the equations are separable, conditions are on constant coordinate curves ($x = 0, x = L, t = 0, r = c$, and $r = 0$), and the equations are linear.

In contrast, an example of a PDE that is *not* separable is the two-dimensional **biharmonic equation** in Cartesian coordinates,

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0 \tag{87}$$

because $u(x, y) = X(x)Y(y)$ gives

$$\frac{X''''}{X} + 2\frac{X''}{X}\frac{Y''}{Y} + \frac{Y''''}{Y} = 0, \tag{88}$$

which cannot be rearranged in the separated form $F(x) = G(y)$ because of the $(X''/X)(Y''/Y)$ term.

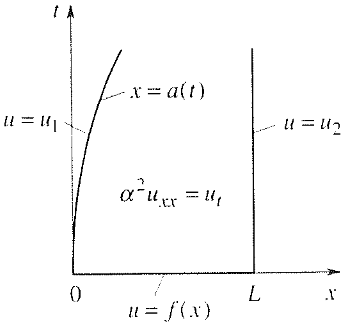


Figure 11. Moving boundary problem.

And as an example where the boundary conditions are not given on constant coordinate curves consider the problem shown in Fig. 11. There, the temperature u_1 is so high that the left end of the rod melts and drips away, so that the boundary condition $u = u_1$ is applied not on the line $x = 0$ but on some curve $x = a(t)$. The latter is an example of a **moving boundary problem**. Generally, such problems defy analytical solution, and we resort to numerical solution techniques – such as the finite difference method that is presented in Section 18.6. An interesting application of the problem shown in Fig. 11 is in the design of a space vehicle that reenters the earth's atmosphere at hypersonic velocity. To protect the craft from the heat thereby generated (remember that meteorites often burn up before reaching the earth's surface) one can design the nose cone to be long enough so that part of it melts away during reentry. A simple one-dimensional model of the heat conduction in the nose cone would be somewhat like the problem shown in Fig. 11.

In this chapter and the next, the sign of the separation constant is always negative, but in Chapter 20, on the Laplace equation, it is negative *or* positive, depending on the specific application.

In subsection 18.3.2 we discuss the rigorous verification of a solution. Normally, our solutions will be only formal, in the sense that such verification will not be carried out.

Finally, in subsection 18.3.3 we show how to use the Sturm–Liouville theory to handle the expansion process that is needed to satisfy the initial condition. Using that theory we are not limited to half- and quarter-range cosine and sine expansions, as we were in subsection 18.3.1. In fact, the half- and quarter-range cases are but special cases of the Sturm–Liouville theory. Remember: Liouville, not Louisville.

Computer software. We can use the *Maple* **sum** command to obtain the infinite series solutions that we generate by separation of variables. As a first illustration, consider the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots, \tag{89}$$

which has the sum $\ln 2$ because it is the Taylor series of $\ln(1+x)$ about $x = 1$. To sum the series, enter

$$\text{sum}\,((-1)\hat{}(n+1)/n, n = 1..\text{infinity}\ ); \tag{90}$$

and return. The result is $\ln 2$. Remarkably, the software not only gives the numerical value, it gives that value in the closed form $\ln 2$ rather than the open form $0.693147\ldots$.

Next, consider the solution (30) given in Example 1, namely,

$$u(x,t) = \frac{400}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin\frac{n\pi x}{10}\, e^{-0.1125n^2 t}. \tag{91}$$

To illustrate, let us use the sum command to sum this series at $x = 1$ and $t = 0.2$. First, change the dummy summation index from n to $2n - 1$ so that the new index n runs continuously from 1 to infinity:

$$u(x,t) = \frac{400}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin\frac{(2n-1)\pi x}{10}\, e^{-0.1125(2n-1)^2 t}. \tag{92}$$

If, to compute $u(1, 0.2)$, we enter

$$\begin{aligned} &\text{sum}\,(400/(Pi * (2 * n - 1)) * \sin((2 * n - 1) * \text{Pi}/10) \\ &* \exp(-0.1125 * (2 * n - 1)\hat{}2 * 0.2), n = 1..\text{infinity}); \end{aligned} \tag{93}$$

the computer is unable to obtain the result in closed form and merely prints the series itself, in a form similar to the right-hand side of (92). Thus, we rerun the command (93) with "infinity" changed to a finite number such as 20. The result, 86.13830281, does not change if we increase the upper summation limit from 20 to 30, say. Thus, it is reasonable to assume that the answer $u(1, 0.2) = 86.13830281$ is correct to that many significant figures. Of course, in a practical sense it would be foolish to insist on 10 significant figure accuracy in a problem like this if $\alpha^2 = 0.0034$ and the other data are known only to two or three significant figures.

If we run the sum command from $n = 1$ to $n = 20$, say, and the result is merely a printout of the 20 terms rather than their numerical sum, we can use the evalf(") command to evaluate their sum.

EXERCISES 18.3

1. Verify by direct substitution that (14) does indeed satisfy the diffusion equation (1a).

2. (*On the sign of separation constant*) Writing the separation constant as $-\kappa^2$ in (6), rather than as κ^2, worked out well, and we also give arguments in Comment 3 as to why the separation constant should be negative. In this exercise we explore that point further.

(a) Show that if we use κ^2 in (6), instead of $-\kappa^2$, then we eventually arrive at the same solution as before, given by (22), although it will be a bit more challenging because complex quantities will arise. HINT: Recall the identity $\sin ix = i \sinh x$ (or, $\sinh ix = i \sin x$). You may use the fact that the only roots of the equation $\sin z = 0$, in the complex plane, are the points $z = n\pi$ on the real axis.
(b) Show that if we use $-\kappa^2$ in (6), as we did, then the relevant Sturm–Liouville problem is given by (53). Show that it follows from Theorem 17.7.2 that κ^2 *must* be nonnegative.

3. In Comment 6 at the end of Example 1 we note that although we began by seeking $u(x,t)$ in product form, the final solution (22) is not itself of that form. Give conditions on u_1, u_2, and $f(x)$ in (1) such that the final solution *will* be of product form (so that its graph does not change its shape, with time, though its magnitude may change with time).

4. (*Separation*) Seeking a solution $u(x,t) = X(x)T(t)$ for the given PDE, carry out steps analogous to equations (3)–(6), and derive ODE's analogous to (7a,b). Take the separation constant to be $-\kappa^2$, as we do in (6). Obtain general solutions of those ODE's (distinguishing any special κ values, as necessary) and use superposition to obtain a solution analogous to the solution (13) of (1a). If the PDE cannot be separated, state that.

(a) $u_{xx} = u_t + 3u$
(b) $u_{xx} + 2u_x = u_t$ HINT: In this case you should find that the value of κ that needs to be distinguished [as we distinguished the case $\kappa = 0$ in (9) and (10)] is $\kappa = 1$, not $\kappa = 0$.
(c) $u_{xx} + 2u_{xt} = u_t$
(d) $u_{xx} + 2u_{xt} = u_{tt}$

5. Can we use superposition to conclude from (9) and (10) that

$$X = A\cos\kappa x + B\sin\kappa x + D + Ex,$$
$$T = Fe^{-\kappa^2\alpha^2 t} + G,$$

and

$$u(x,t) = (A\cos\kappa x + B\sin\kappa x + D + Ex)(Fe^{-\kappa^2\alpha^2 t} + G)?$$

Explain.

6. (*Continuation of Examples 1 and 2*) In each case solve (1), with the boundary conditions (1b) changed as indicated, and for the specified $f(x)$. Use a half- or quarter-range cosine or sine expansion, as appropriate. Evaluate the expansion coefficients explicitly, rather than leaving them in integral form. Also, identify the steady-state solution $u_s(x)$.

(a) $u(0,t) = 20,\ u_x(\pi,t) = 3,$ (i.e., $L = \pi$), $f(x) = 0$
(b) $u(0,t) = 10,\ u_x(2,t) = -5,\ f(x) = 10$
(c) $u(0,t) = 0,\ u_x(2,t) = 0,\ f(x) = 50\sin(\pi x/2)$
(d) $u(0,t) = 0,\ u_x(2,t) = 0,$
$f(x) = 5\sin(\pi x/4) - 12\sin(5\pi x/4)$
(e) $u(0,t) = 25,\ u_x(4,t) = 0,\ f(x) = 25$
(f) $u(0,t) = 25,\ u_x(2,t) = 0,\ f(x) = 0$ for $0 < x < 1$, $f(x) = 25$ for $1 < x < 2$
(g) $u_x(0,t) = u_x(\pi,t) = 0,\ f(x) = 300$
(h) $u_x(0,t) = u_x(3\pi,t) = 0,\ f(x) = 0$ for $0 < x < 2\pi$, $f(x) = 60$ for $2\pi < x < 3\pi$
(i) $u_x(0,t) = u_x(10,t) = 5,\ f(x) = 45 + 5x$
(j) $u_x(0,t) = u_x(5,t) = 3,\ f(x) = 2x$
(k) $u(0,t) = 0,\ u(5,t) = 0,$
$f(x) = \sin\pi x - 37\sin(\pi x/5) + 6\sin(9\pi x/5)$
(l) $u(0,t) = 0,\ u(10,t) = 100,\ f(x) = 0$
(m) $u_x(0,t) = 2,\ u(6,t) = 12,\ f(x) = 0$
(n) $u_x(0,t) = 0,\ u(6,t) = 0,\ f(x) = \sin x$

7. Use (45) and (47) to compute $u(0,t)$, and plot it versus t. At the least, take t = 1,000, 2,000, 5,000, 10,000, 15,000, 20,000 and 30,000. Recall that $\alpha^2 = 0.0034$ and $L = 10$.

8. We stated in Comment 8 that it can be seen by inspection that $K_1 = 40$ and that all the other K_n's are zero. Alternatively, obtain that same resulting by working out the integral

$$K_n = \frac{2}{L}\int_0^L \left(40\sin\frac{\pi x}{L}\right)\sin\frac{n\pi x}{L}\,dx.$$

9. The temperature distribution $u(x,t)$ in a 2-m long brass rod is governed by the problem

$$\alpha^2 u_{xx} = u_t, \qquad (0 < x < 2,\ 0 < t < \infty)$$
$$u(0,t) = u(2,t) = 0, \qquad (t > 0)$$
$$u(x,0) = \begin{cases} 50x, & (0 < x < 1) \\ 100 - 5x, & (1 < x < 2) \end{cases}$$

where $\alpha^2 = 2.9 \times 10^{-5}$ m^2/sec.

(a) Determine the solution for $u(x,t)$.
(b) Compute the temperature at the midpoint of the rod at the end of 1 hour.
(c) Compute the time it will take for the temperature at that point to diminish to 5° C.
(d) Compute the time it will take for the temperature at that point to diminish to 1°C.

10. (*Steady-state solution*) If the solution $u(x,t)$ tends to a steady-state solution $u_s(x)$ as $t \to \infty$, then we can determine $u_s(x)$ from $u(x,t)$ as

$$u_s(x) = \lim_{t\to\infty} u(x,t). \tag{10.1}$$

However, if we are interested *only* in $u_s(x)$ then it is wasteful to first solve for $u(x,t)$. To solve for $u_s(x)$ directly, merely set $u_t = 0$ in the PDE, which step reduces the PDE to an ODE on $u_s(x)$. Solve that ODE subject to the boundary conditions (which, we assume here, do not vary with t). In Example 1, for instance, $u_s(x)$ is governed by the problem

$$\begin{aligned} &\alpha^2 u_s'' = 0, \qquad (0 < x < L) \\ &u_s(0) = u_1, \quad u_s(L) = u_2, \end{aligned}$$

which boundary-value problem is readily solved, its solution being $u_s(x) = u_1 + (u_2 - u_1)x/L$, as obtained in Example 1 by letting $t \to \infty$ in $u(x,t)$. Use this method to find $u_s(x)$ in each case; u_1, u_2, Q_1, Q_2, V, H are constants, and the initial condition is $u(x,0) = f(x)$.

(a) $\alpha^2 u_{xx} = u_t, \;\; u(0,t) = u_1, \;\; u_x(L,t) = Q_2$
(b) $\alpha^2 u_{xx} = u_t, \;\; u_x(0,t) = Q_1, \;\; u(L,t) = u_2$
(c) $\alpha^2 u_{xx} = u_t, \;\; u_x(0,t) = Q_1, \;\; u_x(L,t) = Q_2$ HINT: Show that $u_s(x)$ does not exist if $Q_2 \neq Q_1$, and explain why that result makes sense in physical terms. Show that if $Q_2 = Q_1 \equiv Q$, however, then $u_s(x)$ does exist but contains an undetermined constant, say C. To determine C, integrate the PDE on x, from 0 to L:

$$\alpha^2 \int_0^L u_{xx}\, dx = \int_0^L u_t\, dx, \tag{10.2}$$

and show that

$$\frac{d}{dt}\int_0^L u(x,t)\, dx = 0, \tag{10.3}$$

so that we have the conservation principle

$$\int_0^L u(x,t)\, dx = \text{constant}. \tag{10.4}$$

Use (10.4) to solve for C, thus completing the solution for $u_s(x)$.

(d) $\alpha^2 u_{xx} = u_t + Hu, \;\; u(0,t) = u_1, \;\; u(L,t) = u_2$
(e) $\alpha^2 u_{xx} = u_t + Hu, \;\; u_x(0,t) = Q_1, \;\; u(L,t) = u_2$
(f) $\alpha^2 u_{xx} = u_t + Hu, \;\; u(0,t) = u_1, \;\; u_x(L,t) = Q_2$
(g) $\alpha^2 u_{xx} = u_t + Hu, \;\; u_x(0,t) = Q_1, \;\; u_x(L,t) = Q_2$
(h) $\alpha^2 u_{xx} - Vu_x = u_t, \;\; u(0,t) = u_1, \;\; u(L,t) = u_2$
(i) $\alpha^2 u_{xx} - Vu_x = u_t, \;\; u_x(0,t) = Q_1, \;\; u(L,t) = u_2$
(j) $\alpha^2 u_{xx} - Vu_x = u_t, \;\; u(0,t) + 5u_x(0,t) = 3, \;\; u(L,t) = 10$
(k) $\alpha^2 u_{xx} - Vu_x = u_t, \;\; u_x(0,t) = 0,$
$u(L,t) + 2u_x(L,t) = -5$

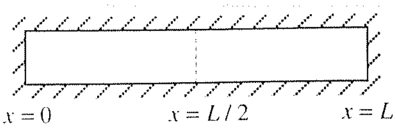
11. (*Existence of steady state*) For the problem

$$\begin{aligned} &\alpha^2 u_{xx} = u_t + F(x), \qquad (0 < x < L, \; 0 < t < \infty) \\ &u_x(0,t) = Q_1, \quad u_x(L,t) = Q_2, \quad u(x,0) = f(x), \end{aligned}$$

show that a steady state does not exist unless a certain condition is satisfied by Q_1, Q_2, and $F(x)$. Assuming that that condition is satisfied, solve for $u_s(x)$.

12. (*Steady-state extrusion*) In Section 18.2 we derive equation (20) governing the temperature distribution $u(x,t)$ in a heated rod being drawn through an extrusion die, as sketched there in Fig. 2. Actually, (20) holds both inside the furnace ($x < 0$) and outside the furnace ($x > 0$), but with different h's and u_∞'s. Let $h = h_f$ and h_a, and let $u_\infty = u_f$ and u_a inside and outside the furnace, respectively. Assuming steady-state operation so that $u = u_s(x)$, propose a suitable set of boundary conditions and solve for $u_s(L)$, the rod temperature at the die. HINT: It will be a helpful approximation to consider the rod to extend from $-\infty$ to $+\infty$. Over $-\infty < x < 0$ use h_f and u_f in the ODE, and over $0 < x < \infty$ use h_a and u_a. Solve over $x < 0$ and $x > 0$ separately, and apply suitable boundary conditions at $x = -\infty$ and $x = +\infty$, as well as suitable "matching conditions" at $x = 0$.

13. (*Diffusion of one gas into another*) Consider a cylindrical compressed-gas container of length L, divided in half by a baffle (see sketch). To the left of the baffle is a gas of species A, and to the right of it is a different gas of species B. Suppose they are at the same pressure, so that when the baffle is removed at time $t = 0$ the two gases proceed to mix by diffusion



alone. Considering species A, say, its concentration $c_A(x,t)$ moles/cm^3 is governed by the problem

$$D\frac{\partial^2 c_A}{\partial x^2} = \frac{\partial c_A}{\partial t}, \qquad (0 < x < L,\ 0 < t < \infty) \qquad (13.1)$$

$$\frac{\partial c_A}{\partial x}(0,t) = \frac{\partial c_A}{\partial x}(L,t) = 0, \qquad (0 < t < \infty) \qquad (13.2)$$

$$c_A(x,0) = \begin{cases} c_0 & 0 < x < L/2 \\ 0, & L/2 < x < L \end{cases} \qquad (13.3)$$

where D is the diffusion coefficient and D and c_0 are constants.

(a) Solve for $c_A(x,t)$. From $c_A(x,t)$ determine the steady-state solution

$$c_{As}(x) = \lim_{t\to\infty} c_A(x,t). \qquad (13.4)$$

Draw a neat sketch of what you expect the graph of $c_A(x,t)$ versus x to look like, at $t = 0$, $t = \infty$ and at a couple of intermediate times. Label any key values and features of those graphs.

(b) Integrating equation (13.1) with respect to x, from 0 to L, show that

$$\int_0^L c_A(x,t)\,dx = \text{constant}, \qquad (13.5)$$

which says that the total amount of A is conserved. This result makes sense physically since the container is sealed [note (13.2)] and the gas is being neither created nor destroyed. (The point here is that it may be possible to learn something about the solution, from the PDE and boundary and initial conditions, without actually obtaining the full solution.)

(c) Solve for $c_{As}(x)$ directly, i.e., by solving

$$Dc''_{As}(x) = 0; \quad c'_{As}(0) = c'_{As}(L) = 0$$

and using equation (13.5). Your result should, of course, be the same as in part (a).

14. (*Conduction in metal ring*) Consider the conduction of heat in a circular metal ring, the surface of which is insulated. Actually, whether the shape is a circle or an ellipse or whatever is irrelevant insofar as the heat conduction is concerned. If we measure x along the ring, from some starting point, and denote the length of the ring as L, then the temperature $u(x,t)$ is defined on $-\infty < x < \infty$ and is an L-periodic function of x.

(a) Solve the heat equation $\alpha^2 u_{xx} = u_t$ $(-\infty < x < \infty,\ 0 < t < \infty)$ subject to the initial condition $u(x,0) = f(x)$, where $f(x)$ is L-periodic on $-\infty < x < \infty$. Letting $t \to \infty$ in your solution, show that the steady-state solution is a constant temperature that is the average value of f. HINT: Since u is L-periodic in x, it must be expressible in the Fourier series form

$$u(x,t) = a_0(t) + \sum_{n=1}^{\infty}\left[a_n(t)\cos\frac{2n\pi x}{L} + b_n(t)\sin\frac{2n\pi x}{L}\right], \qquad (14.1)$$

where a_0, a_n, b_n vary with time. Putting (14.1) into the PDE and matching the coefficients of 1, $\cos(2n\pi x/L)$, and $\sin(2n\pi x/L)$ on the left- and right-hand sides, derive simple ODE's governing $a_0(t), a_n(t), b_n(t)$. Solve these ODE's and then apply the initial condition $u(x,0) = f(x)$. Give integral formulas for the evaluation of any constants. (Exercise 29 re-examines this problem using the Sturm–Liouville theory.)

(b) Integrating $\alpha^2 u_{xx} = u_t$ from 0 to L, derive the conservation principle

$$\int_0^L u(x,t)\,dx = \text{constant}. \qquad (14.2)$$

(c) Derive the steady-state solution $u_s(x)$ again, this time by solving

$$\alpha^2 u''_s(x) = 0, \qquad (0 < x < L)$$
$$u_s(0) = u_s(L), \quad u'_s(0) = u'_s(L),$$

and using (14.3). Your result should be the same as found in part (a).

15. (*Presence of a constant source term*) Consider the problem

$$\begin{aligned} &\alpha^2 u_{xx} = u_t - F, && (0 < x < L,\ 0 < t < \infty) \\ &u(0,t) = 0, \quad u(L,t) = 50, && (0 < t < \infty) \\ &u(x,0) = f(x), && (0 < x < L) \end{aligned} \qquad (15.1,2,3)$$

where the source term F is assumed to be a constant. Solve for $u(x,t)$. Expansion coefficients may be left in integral form. HINT: This is the first problem in which the PDE is *nonhomogeneous*, namely, $L[u] = \alpha^2 u_{xx} - u_t = -F$. Observe that if we seek $u(x,t) = X(x)T(t)$ as in the text examples, and attempt to carry out the separation process, we obtain

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T} - \frac{F}{\alpha^2}\frac{1}{XT}. \qquad (15.4)$$

Because of the last term in (15.4), which contains both x

and t dependence, we are unable to complete the separation process successfully. That is, we are unable to get all of the x dependence on one side of the equation and all of the t dependence on the other side. Thus, we suggest seeking u in the form

$$u(x,t) = u_s(x) + X(x)T(t) \tag{15.5}$$

instead, where $u_s(x)$ is the steady-state solution. In steady state, $u_t \to 0$ and $u(x,t) \to u_s(x)$, so (15.1) and (15.2) give

$$\alpha^2 u_s''(x) = -F; \qquad u_s(0) = 0, \quad u_s(L) = 50. \tag{15.6}$$

Putting (15.5) into (15.1), show that the u_s term cancels the troublesome F term, so that this time the separation can be successfully completed to yield the familiar result

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T} = \text{constant} = -\kappa^2. \tag{15.7}$$

Then solve (15.6) for u_s and (15.7) for X and T, and impose the conditions (15.2) and (15.3) on $u(x,t) = u_s(x) + X(x)T(t)$. Regarding the form of (15.5), observe that, in *physical* terms, u_s is the steady-state solution and XT is a transient quantity needed to match the initial condition $u(x,0) = f(x)$ with the steady-state solution $u_s(x)$. In *mathematical* terms, u_s is a particular solution (i.e., a solution of the full equation $\alpha^2 u_{xx} - u_t = -F$) and XT is a solution of the associated homogeneous equation $\alpha^2 u_{xx} - u_t = 0$.

16. Repeat Exercise 15 with (15.2) and (15.3) changed as follows.

(a) $u(0,t) = u(L,t) = u(x,0) = 0$
(b) $u_x(0,t) = u(L,t) = u(x,0) = 0$
(c) $u(0,t) = u_x(L,t) = u(x,0) = 0$
(d) $u(0,t) = 0, u_x(L,t) = -20, u(x,0) = 0$

17. (*Presence of nonconstant source term*) In Exercise 15 we include a source term F that is a constant, although the solution method outlined therein would have worked even if F were a nonconstant function of x. In this exercise we consider the problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t - F(x,t), \qquad && (0 < x < L,\ 0 < t < \infty) \\ u(0,t) &= u(L,t) = 0, && (0 < t < \infty) \\ u(x,0) &= 0, && (0 < x < L) \end{aligned} \tag{17.1}$$

where the source term F is allowed to be a function of x and t. To solve, we can use essentially the same **eigenvector expansion method** that we use in Section 11.3.2 to solve the nonhomogeneous matrix problem $\mathbf{A}\mathbf{x}{=}\Lambda\mathbf{x} + \mathbf{c}$. [In this case the matrix operator $\mathbf{A}$ is analogous to the partial differential operator $\alpha^2\partial^2/\partial x^2 - \partial/\partial t$, $\mathbf{c}$ is analogous to $-F(x,t)$, and $\Lambda = 0$.] There, we expand $\mathbf{x}$ and $\mathbf{c}$ in terms of the basis provided by the eigenvectors of $\mathbf{A}$. In the present case that step amounts to expanding $u(x,t)$ and $F(x,t)$ in terms of the $\sin\frac{n\pi x}{L}$ eigenfunctions provided by the relevant Sturm-Liouville problem. Thus, the problem that we pose is as follows.

(a) Solve for $u(x,t)$ by seeking

$$u(x,t) = \sum_{n=1}^{\infty} g_n(t) \sin\frac{n\pi x}{L} \tag{17.2}$$

and expanding

$$F(x,t) = \sum_{n=1}^{\infty} F_n(t) \sin\frac{n\pi x}{L}, \tag{17.3}$$

where the

$$F_n(t) = \frac{2}{L}\int_0^L F(x,t)\sin\frac{n\pi x}{L}\,dx \tag{17.4}$$

coefficients are considered as known (i.e., computable) functions of t. Thus, show that

$$u(x,t) = \sum_{n=1}^{\infty}\left[\int_0^t F_n(\tau)e^{(n\pi\alpha/L)^2(\tau - t)}\,d\tau\right]\sin\frac{n\pi x}{L}. \tag{17.5}$$

(b) Work out the solution (17.5) for the case where $F(x,t) = e^{-t}$.
(c) Modify the solution procedure described in part (a) if the left end condition is changed from $u(0,t) = 0$ to $u_x(0,t) = 0$.

18. (*Superposition*) Show that the solution to the problem

$$\begin{aligned} &\alpha^2 u_{xx} = u_t + g(x,t), \qquad (0 < x < L,\ 0 < t < \infty) \\ &u(0,t) = p(t), \quad u(L,t) = q(t), \quad u(x,0) = f(x) \end{aligned} \tag{18.1}$$

can be expressed as $u = u_1 + u_2 + u_3 + u_4$, where $u_1, \ldots, u_4$ are solutions of the four problems

$$\begin{aligned} &\alpha^2 u_{1xx} = u_{1t} + g(x,t), \\ &u_1(0,t) = u_1(L,t) = u_1(x,0) = 0, \end{aligned}$$

$$\begin{aligned} &\alpha^2 u_{2xx} = u_{2t}, \\ &u_2(0,t) = p(t), \quad u_2(L,t) = u_2(x,0) = 0, \end{aligned}$$

$$\alpha^2 u_{3xx} = u_{3t},$$

$$u_3(0,t) = 0, \quad u_3(L,t) = q(t), \quad u_3(x,0) = 0,$$

$$\begin{aligned} &\alpha^2 u_{4xx} = u_{4t}, \\ &u_4(0,t) = u_4(L,t) = 0, \quad u_4(x,0) = f(x), \end{aligned}$$

each on the domain $0 < x < L, \ \ 0 < t < \infty$.

19. (*Nonhomogeneous Neumann conditions*) Solve the conduction problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t, \qquad (0 < x < L, \ \ 0 < t < \infty) \\ u_x(0,t) &= -1, \quad u_x(L,t) = 0, \quad u(x,0) = 0 \end{aligned} \tag{19.1}$$

for $u(x,t)$. HINT: You should find that the standard separation-of-variables procedure has difficulty coping with the boundary conditions if $u_x(0,t) \neq u_x(L,t)$, as is true in this case. To proceed successfully, we suggest that you change from $u(x,t)$ to $v(x,t)$ according to

$$u(x,t) = \frac{(x-L)^2}{2L} + v(x,t).$$

Then, show that v can be split by superposition into $v = v_1 + v_2$, where

$$\begin{aligned} \alpha^2 v_{1xx} &= v_{1t}, \qquad (0 < x < L, \ \ 0 < t < \infty) \\ v_{1x}(0,t) &= v_{1x}(L,t) = 0, \quad v_1(x,0) = -\frac{(x-L)^2}{2L} \end{aligned}$$

and

$$\begin{aligned} \alpha^2 v_{2xx} &= v_{2t} - \frac{\alpha^2}{L}, \qquad (0 < x < L, \ \ 0 < t < \infty) \\ v_{2x}(0,t) &= v_{2x}(L,t) = v_2(x,0) = 0. \end{aligned}$$

The solution for v_2 can be found easily as a function of t alone. Solve for v_2. (But you need not solve for v_1.)

20. (*Variable end conditions*) Thus far, our Dirichlet-type boundary conditions have been of constant type, e.g., $u(0,t) = 50$. Here, we consider nonconstant conditions. Consider the problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t, \qquad (0 < x < L, \ \ 0 < t < \infty) \\ u(0,t) &= p(t), \quad u(L,t) = q(t), \quad u(x,0) = f(x), \end{aligned} \tag{20.1}$$

where $p(t), q(t)$, and $f(x)$ are prescribed. Changing dependent variables from $u(x,t)$ to $v(x,t)$ according to

$$u(x,t) = v(x,t) + \left(1 - \frac{x}{L}\right) p(t) + \frac{x}{L}\, q(t), \tag{20.2}$$

show that the problem governing v is precisely of the type treated in Exercise 17. NOTE: Observe how an "input" can be moved from the boundary conditions to the PDE. In the present case, the PDE on u was homogeneous and the boundary conditions were nonhomogeneous; following the change of variables (20.2), you should find that the PDE on v is nonhomogeneous and the boundary conditions are homogeneous.

21. (*Newton cooling*) Consider the conduction of heat in a rod, the lateral surface of which is not insulated. If heat is convected from the rod to the environment, the PDE governing the temperature $u(x,t)$ is

$$\alpha^2 u_{xx} = u_t + h(u - u_\infty), \tag{21.1}$$

where the constants h and u_∞ are the convective heat transfer coefficient and the ambient temperature, respectively. Our interest in (21.1) lies in the Newton cooling term $h(u - u_\infty)$. Although it is not essential, one normally begins by eliminating the u_∞ term by setting $v(x,t) = u(x,t) - u_\infty$ and considering, instead, $\alpha^2 v_{xx} = v_t + hv$.

(a) Solve the Newton cooling problem

$$\begin{aligned} \alpha^2 v_{xx} &= v_t + hv, \qquad (0 < x < L, \ \ 0 < t < \infty) \\ v(0,t) &= v(L,t) = 50 \quad v(x,0) = f(x) \end{aligned} \tag{21.2}$$

by separation of variables, leaving expansion coefficients in integral form. HINT: Seeking $v(x,t) = v_s(x) + X(x)T(t)$, where $v_s(x)$ is the steady-state temperature distribution, show that

$$\frac{X''}{X} = \frac{1}{\alpha^2}\,\frac{T' + hT}{T} = \text{constant} = -\kappa^2. \tag{21.3}$$

(b) Solve (21.2) by omitting the $v_s(x)$ term and seeking $v(x,t) = X(x)T(t)$. That is, show that the inclusion of the $v_s(x)$ term in the solution form is not essential.

22. Show that the change of variables

$$w(x,t) = e^{ht}[u(x,t) - u_\infty]$$

reduces (21.1), above, to the simpler and more familiar form

$$\alpha^2 w_{xx} = w_t.$$

23. (*Flow of electricity in a cable*) The voltage $v(x,t)$ (volts) and the current $I(x,t)$ (amperes) in a long underground insulated cable are governed by the PDE's

$$\begin{aligned} v_{xx} &= LCv_{tt} + (rC + Lg)v_t + rgv, \\ I_{xx} &= LCI_{tt} + (rC + Lg)I_t + rgI, \end{aligned} \tag{23.1,2}$$

where L, C, r, g are positive constants: L is the inductance (henries/kilometer), C is the capacitance to ground (farads/kilometer), r is the resistance (ohms/kilometer), and g is the leakance to ground (mhos/kilometer). These PDE's are called the **telephone equations** and are seen to be of wave (hyperbolic) type. Often, as in telegraph transmission, L and g can be neglected, in which case (23.1) and (23.2) reduce to the **telegraph equations**

$$v_{xx} = rCv_t,$$
$$I_{xx} = rCI_t,$$

which are of diffusion type. In the present example we suppose only that $L \approx 0$. Considering only the voltage v, we then have

$$v_{xx} = rCv_t + rgv. \tag{23.3}$$

[Comparing (23.3) with the PDE in Exercise 21(a), we see that the two phenomena are analogous, with the lateral heat loss to the environment corresponding to the voltage loss due to leakage to the ground.] Suppose that the line is of length L, the voltage at $x = 0$ is maintained (for a "long time") at 12 volts, the voltage at $x = L$ is maintained (for a "long time") at 6 volts, and then, beginning at $t = 0$, the left end is grounded. Thus,

$$\begin{aligned} v(0,t) &= 0, \quad (0 < t < \infty) \\ v(L,t) &= 6. \quad (0 < t < \infty) \end{aligned} \tag{23.4,5}$$

The initial condition $v(x, 0)$ is not given but can be deduced from (23.3) together with the information that the ends have been maintained at 12 and 6 volts, respectively, for a long time.

(a) Determine $v(x, 0)$.
(b) Determine the steady-state solution $v_s(x)$.
(c) Solve for $v(x,t)$ by separation of variables. Fourier expansion coefficients may be left in integral form. HINT: As in Exercise 21, it will be most convenient to seek $v(x,t) = v_s(x) + X(x)T(t)$.

24. (*Conduction in a sphere*) Consider the radial conduction of heat within a solid sphere. If the temperature u is a function only of the spherical polar coordinate ρ and the time t, then $\alpha^2\nabla^2 u = u_t$ becomes

$$\alpha^2\left(u_{\rho\rho} + \frac{2}{\rho}u_\rho\right) = u_t. \tag{24.1}$$

(a) Setting $u(\rho, t) = v(\rho, t)/\rho$, show that v needs to satisfy the more familiar PDE

$$\alpha^2 v_{\rho\rho} = v_t. \tag{24.2}$$

(b) Use the idea contained in part (a) to solve the problem

$$\alpha^2\left(u_{\rho\rho} + \frac{2}{\rho}u_\rho\right) = u_t, \qquad (0 < \rho < a,\ 0 < t < \infty)$$
$$u(a,t) = 0, \quad (0 < t < \infty)$$
$$u(\rho, 0) = f(\rho), \quad (0 < \rho < a)$$

where $u(0,t)$ is bounded. Expansion coefficients may be left in integral form. (Exercise 28 reexamines this problem, using the Sturm–Liouville theory.)

EXERCISES FOR THE OPTIONAL SECTIONS 18.3.2, 18.3.3

25. (*Uniqueness*) (a) Consider the problem

$$\begin{aligned} &\alpha^2 u_{xx} = u_t + f(x,t), \qquad (0 < x < L,\ 0 < t < T) \\ &u(0,t) = p(t), \quad u(L,t) = q(t), \quad u(x,0) = r(x), \end{aligned} \tag{25.1,2}$$

where T is arbitrarily large. To establish the *uniqueness* of the solution, suppose that $u_1(x,t)$ and $u_2(x,t)$ are two solutions, and define

$$w(x,t) = u_1(x,t) - u_2(x,t). \tag{25.3}$$

Show that w satisfies the "homogenized" problem

$$\alpha^2 w_{xx} = w_t, \qquad (0 < x < L,\ 0 < t < T)$$
$$w(0,t) = 0, \quad w(L,t) = 0, \quad w(x,0) = 0.$$

Proceeding formally, show that

$$\frac{d}{dt}\int_0^L w^2\,dx = 2\int_0^L ww_t\,dx = 2\alpha^2\int_0^L ww_{xx}\,dx. \tag{25.4}$$

Integrating the last integral by parts and then integrating both sides of the equation on t from 0 to t, show that

$$\int_0^L w^2(x,t)\,dx = -\alpha^2\int_0^t\int_0^L w_x^2(x,\tau)\,dx\,d\tau. \tag{25.5}$$

Explain why it follows from (25.5) that $w(x,t) = 0$ throughout $0 < x < L$, $0 < t < T$. Thus, it must be true that $u_1(x,t) = u_2(x,t)$ for any solutions u_1 and u_2, so that the solution to (25.1) and (25.2) must be unique.
(b) Repeat part (a), but change $u(L,t) = q(t)$ in equation (25.2) to $u_x(L,t) = q(t)$.

(c) Establish the uniqueness of the solution to the problem

$$\begin{gathered}\alpha^2 u_{xx} = u_t + f(x,t), \qquad (0 < x < L,\ 0 < t < T)\\ u(0,t) = p(t), \quad u(L,t) + \beta u_x(L,t) = q(t),\\ u(x,0) = r(x),\end{gathered}$$

where T is arbitrarily large and $\beta > 0$, following essentially the same lines in part (a).

26. With M_n defined by (50), use the ratio test to show that $\sum_{n=1}^{\infty} M_n$ converges, as we claimed.

27. (a)–(n) The problems in Exercise 6 are to be solved using half- or quarter-range expansions. Solve the corresponding problem in Exercise 6 again, this time using the Sturm–Liouville theory.

28. (*Conduction in a sphere*) In Exercise 24(b) we suggest using the change of variables $u(\rho,t) = v(\rho,t)/\rho$ to reduce the PDE to the form $\alpha^2 v_{\rho\rho} = v_t$ that is studied in this section. Here we ask you to solve the problem in Exercise 24(b) directly by seeking $u(\rho,t) = R(\rho)T(t)$ and using separation of variables. HINT: The ODE governing $R(\rho)$ can be solved in terms of Bessel functions using the formulas given in Section 4.6.6, and these results can be simplified using the formulas

$$J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\sin x, \qquad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\,\cos x.$$

You will need to distinguish the cases $\kappa \neq 0$ and $\kappa = 0$.

29. (*Conduction in a metal ring*) Here we reconsider the problem of Exercise 14(a). There we consider the x domain to be $-\infty < x < \infty$, so $u(x,t)$ and $f(x)$ were L-periodic in x and could be expanded in Fourier series. Alternatively, think of the x domain as finite: $0 < x < L$. That step creates the boundaries $x = 0$ and $x = L$, so we need boundary conditions there. Although we don't know u or u_x there, we do know that, physically, the ends $x = 0$ and $x = L$ are abutting, so both the temperature u and the heat flux (proportional to u_x) must be continuous there. Thus, we can pose the problem (on the *finite* interval $0 < x < L$) as

$$\begin{gathered}\alpha^2 u_{xx} = u_t, \qquad (0 < x < L,\ 0 < t < \infty)\\ u(0,t) = u(L,t), \quad u_x(0,t) = u_x(L,t), \quad u(x,0) = f(x).\end{gathered}$$

Solve that problem by separation of variables, leaving expansion coefficients in integral form. HINT: The Sturm–Liouville problem that arises on $X(x)$ will have periodic boundary conditions.

30. (a) Solve the problem

$$\begin{gathered}u_{xx} - 2u_x = u_t, \qquad (0 < x < L,\ 0 < t < \infty)\\ u(0,t) = u(L,t) = 50, \quad u(x,0) = 0\end{gathered} \qquad (30.1)$$

by separation of variables. HINT: Seeking $u(x,t) = X(x)T(t)$, obtain

$$\begin{aligned} X'' - 2X' + \kappa^2 X &= 0,\\ T' + \kappa^2 T &= 0.\end{aligned} \qquad (30.2,3)$$

With $X(x) = e^{\lambda x}$, obtain $\lambda = 1 \pm \sqrt{1-\kappa^2}$, so

$$\begin{aligned} X(x) &= e^x\left(C_1 e^{\sqrt{1-\kappa^2}\,x} + C_2 e^{-\sqrt{1-\kappa^2}\,x}\right),\\ T(t) &= C_3 e^{-\kappa^2 t}.\end{aligned} \qquad (30.4,5)$$

However, expecting oscillatory functions (for the eventual expansion that will be needed), anticipate that the κ's will be greater than unity and write $\lambda = 1 \pm i\sqrt{\kappa^2 - 1}$ instead, so

$$X(x) = e^x\left(C_3 \cos\sqrt{\kappa^2-1}\,x + C_2 \sin\sqrt{\kappa^2-1}\,x\right). \qquad (30.6)$$

Distinguish the case $\kappa = 1$ because if $\kappa = 1$ then (30.6) reduces to $X(x) = C_3 e^x$, which falls short of being a general solution of (30.2). We don't *need* to also distinguish the case $\kappa = 0$ because if $\kappa = 0$ then (30.4) and (30.5) do give the general solutions of (30.2) and (30.3), respectively. However, the case $\kappa = 0$ is of special importance because it gives the steady-state part of the solution [since it reduces $T(t)$ to a constant]. Thus, use the "three-tier" solutions

$$X(x) = \begin{cases} e^x\left(A\cos\omega x + B\sin\omega x\right), & \kappa \neq 0,1\\ e^x(C + Dx), & \kappa = 1\\ E + Fe^{2x}, & \kappa = 0\end{cases}$$

$$T(t) = \begin{cases} Ge^{-\kappa^2 t}, & \kappa \neq 0,1\\ He^{-t}, & \kappa = 1\\ I, & \kappa = 0\end{cases}$$

where $\omega \equiv \sqrt{\kappa^2 - 1}$, for brevity, and form $u(x,t)$ as the sum of their respective products.

(b) Same as part (a), but with $u(0,t) = 0$ and $u(L,t) = 50$. Leave expansion coefficients in integral form.

31. (a) Show that

$$\int_0^{z_n} xJ_0(x)\,dx = z_n J_1(z_n) \qquad (31.1)$$

where z_n is any root of $J_0(x) = 0$. HINT: Integrate the

Bessel equation $(xJ_0')' + xJ_0 = 0$ from 0 to z_n and use the relation $J_0'(x) = -J_1(x)$ [from Exercise 4, Section 4.6].

(b) Then, use (31.1) to verify the last step in (83).

18.4 Fourier and Laplace Transforms (Optional)

In Section 18.3 the x domain is always finite, namely, $0 < x < L$. Semi-infinite $(0 < x < \infty)$ and infinite $(-\infty < x < \infty)$ domains are also important, and it is these cases that we now address. We organize our discussion around two examples.

EXAMPLE 1. *Heat Conduction in an Infinite Rod.* Let us begin with the problem

$$\alpha^2 u_{xx} = u_t, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \tag{1a}$$

$$u(x,0) = f(x), \qquad (-\infty < x < \infty) \tag{1b}$$

summarized in Fig. 1.

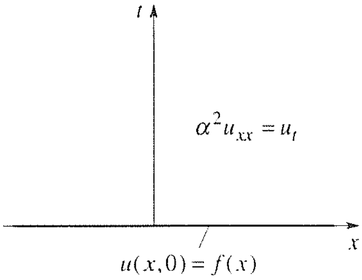


Figure 1. Infinite rod problem.

If we regard the infinite rod as the limiting case of a finite rod, on $-L < x < L$, as $L \to \infty$, and recall that boundary conditions are needed for the finite rod, we might well anticipate that some form of boundary conditions will be needed for the infinite rod at $x = \pm\infty$. But since a suitable form for those boundary conditions may not yet be apparent, let us defer that issue for the moment, in the hope that the solution process itself may provide a clue.

In selecting a solution technique, remember that if we use the method of separation of variables then we need, in the final step, to expand the $f(x)$ in (1b) in a Fourier-type series. If the x domain is finite, then that series will be a half- or quarter-range cosine or sine series, or a generalized Fourier series containing the orthogonal eigenfunctions of a relevant Sturm–Liouville problem. If the x domain is infinite, then we can still use separation of variables, provided that $f(x)$ is periodic with finite period, in which case we can expand $f(x)$ in a classical Fourier series of cosines and sines.

However, in this example we have an infinite x domain and are interested in f's that are *not* periodic, such as $e^{-|x|}$ (Fig. 2a) and the rectangular pulse shown in Fig. 2b. Such f's can be represented not by Fourier series but by Fourier integrals, which fact suggests seeking a solution for u in Fourier integral form or, equivalently and more conveniently, using a Fourier transform.

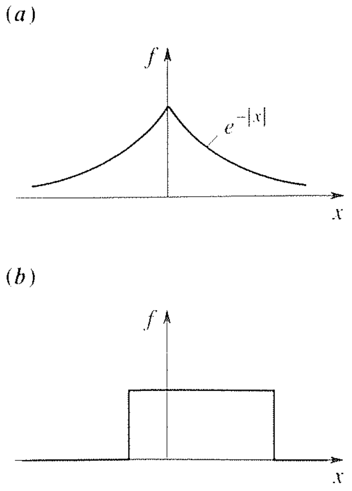


Figure 2. Nonperiodic f's.

Thus, let us Fourier transform (1a) with respect to x:

$$F\{\alpha^2 u_{xx}\} = F\{u_t\}, \tag{2a}$$

$$\alpha^2 F\{u_{xx}\} = \int_{-\infty}^{\infty} \frac{\partial u}{\partial t} e^{-i\omega x}\, dx, \tag{2b}$$

$$\alpha^2 (i\omega)^2 \hat{u} = \frac{d}{dt} \int_{-\infty}^{\infty} u(x,t) e^{-i\omega x}\, dx \tag{2c}$$

$$= \frac{d\hat{u}}{dt}, \tag{2d}$$

so

$$\frac{d\hat{u}}{dt} + \alpha^2\omega^2\hat{u} = 0, \tag{3}$$

with solution

$$\hat{u} = Ae^{-\alpha^2\omega^2 t}. \tag{4}$$

In passing from (2a) to (2b) we used the linearity of $F\{\ \}$ on the left, and the definition of the Fourier transform on the right. Next, we used

$$F\{u_{xx}\} = (i\omega)^2 F\{u\} = (i\omega)^2\hat{u} \tag{5}$$

on the left, and the Leibniz differentiation formula

$$\frac{d}{dt}\int_{-\infty}^{\infty} u(x,t)e^{-i\omega x}\,dx = \int_{-\infty}^{\infty} \frac{\partial u}{\partial t}\,e^{-i\omega x}\,dx \tag{6}$$

on the right. For (5) to hold, we need

$$u \to 0 \text{ and } u_x \to 0 \text{ as } x \to \pm\infty, \tag{7}$$

so let us adopt (7) as our boundary conditions. For the output u (and u_x) to tend to zero as $x \to \pm\infty$, we expect that we will need to restrict the input f to die out sufficiently fast, as $x \to \pm\infty$, as well. However, as suggested in Section 18.3 let us proceed *formally* to a solution without getting bogged down with such technical points. If we like, we can then rigorously verify that that solution does satisfy the PDE and any boundary and initial conditions (if indeed it does).

Notice carefully that by Fourier transforming (1a) with respect to x we convert the *partial* differential equation (1a) on $u(x,t)$ to the *ordinary* differential equation (3) on $\hat{u}(\omega,t)$, in which ω appears as a parameter – that is, as a constant.

To evaluate the integration constant A in (4) we impose the initial condition (1b), but first we must take the Fourier transform of (1b): $\hat{u}(\omega,0) = \hat{f}(\omega)$. Thus,

$$\hat{u}\Big|_{t=0} = \hat{f}(\omega) = \left(Ae^{-\alpha^2\omega^2 t}\right)\Big|_{t=0} = A, \tag{8}$$

so $A = \hat{f}(\omega)$,[*] and (4) becomes

$$\hat{u}(\omega,t) = \hat{f}(\omega)e^{-\alpha^2\omega^2 t}. \tag{9}$$

Finally, using entry 6 of Appendix D and the Fourier convolution property (entry 21), it follows from (9) that

$$u(x,t) = f(x) * \frac{1}{2\alpha\sqrt{\pi t}}\,e^{-x^2/(4\alpha^2 t)} \tag{10}$$

or

$$u(x,t) = \frac{1}{2\alpha\sqrt{\pi t}}\int_{-\infty}^{\infty} f(\xi)e^{-(x-\xi)^2/(4\alpha^2 t)}\,d\xi. \tag{11}$$

[*] You may be concerned that A was to be a constant, yet $\hat{f}(\omega)$ is a function of ω. The point to keep in mind is that (3) is a differential equation in t, and ω is regarded there as a constant.

COMMENT 1. As a general principle it is important to check one's result for any special cases for which the solution is known. The simplest case that comes to mind is the case where

$$f(x) = \text{constant} \equiv F, \tag{12}$$

for in that case it is obvious that the solution to (1) is simply $u(x,t) = F$. In fact, if we set $f(\xi) = F$ in (11) and evaluate the integral we do obtain $u(x,t) = F$ (Exercise 1). It is striking that (11) gives the correct result for the case where $f(x) = F$ because in that case $\hat{f}(\omega)$ does not even exist, and the solution $u(x,t) = F$ violates the assumption that $u \to 0$ as $x \to \pm\infty$.*

COMMENT 2. As a more interesting special case, let

$$f(x) = \begin{cases} F, & x > 0 \\ 0, & x < 0 \end{cases} = FH(x), \tag{13}$$

where F is a constant and $H(x)$ is the Heaviside function. In this case, putting (13) into (11) gives (Exercise 2)

$$u(x,t) = \frac{F}{2}\left[1 + \text{erf}\left(\frac{x}{2\alpha\sqrt{t}}\right)\right], \tag{14}$$

where

$$\boxed{\text{erf}\,(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-\xi^2}\,d\xi} \tag{15}$$

is a tabulated function known as the **error function**. That is, the integral in (15) cannot be evaluated in terms of the so-called elementary functions, so it is given its own name, the "error function," and its properties, tabulated values, and computational formulas for it can be found in the literature.† The factor $2/\sqrt{\pi}$ is included to "normalize" erf (x) so that erf $(\infty) = 1$ because

$$\boxed{\int_0^\infty e^{-\xi^2}\,d\xi = \frac{\sqrt{\pi}}{2},} \tag{16}$$

*The greater-than-expected validity of (11) can be traced to the inversion of the order of integration that is inherent in the Fourier convolution step.

†The trend in computing has been away from tabulations and toward approximate expressions, not only for erf (x) but for the various special functions: Bessel functions, the gamma function, and so on. For instance, the formula

$$\text{erf}\,(x) \approx 1 - (a_1 p + a_2 p^2 + a_3 p^3)e^{-x^2},$$

where

$$p = \frac{1}{1 + 0.47047x}, \quad a_1 = 0.3480242, \quad a_2 = -0.0958798, \quad a_3 = 0.7478556,$$

developed by C. Hastings, Jr., is uniformly accurate, over $0 \le x < \infty$, to ± 0.000025. The latter formula is an example of a common form of approximation known as **rational function approximation** because the function exp $(x^2)[1 - \text{erf}\,(x)]$ is being approximated by a rational function of x, namely, the ratio of two polynomials (as can be seen by combining the terms in $a_1 p + a_2 p^2 + a_3 p^3$ over a common denominator). For approximations such as this, see M. Abramowitz and I. Stegun (eds.), National Bureau of Standards Applied Math Series, 1964, or Y. L. Luke, *Mathematical Functions and Their Approximation* (New York: Academic Press, 1975).

derivation of which result is outlined in Exercise 9 of Section 4.5. The graph of $\operatorname{erf}(x)$ is shown in Fig. 3 for $x > 0$; for $x < 0$ we rely on the fact that the error function is an odd function (Exercise 4), so that $\operatorname{erf}(-x) = -\operatorname{erf}(x)$. Besides the integral from 0 to x, in (15), we also encounter the integral from x to ∞ so frequently that we also define a **complementary error function**,

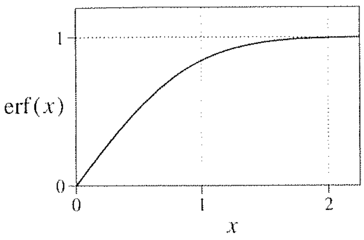


Figure 3. The error function.

$$\begin{aligned}\operatorname{erfc}(x) &= \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-\xi^2}\,d\xi = \operatorname{erf}(\infty) - \operatorname{erf}(x)\\ &= 1 - \operatorname{erf}(x).\end{aligned} \tag{17}$$

The solution (14) is plotted in Fig. 4 at representative times. Observe carefully how the initially discontinuous temperature distribution smooths out as t increases. This result illustrates the fact that, in physical terms, diffusion is a *smoothing process*.

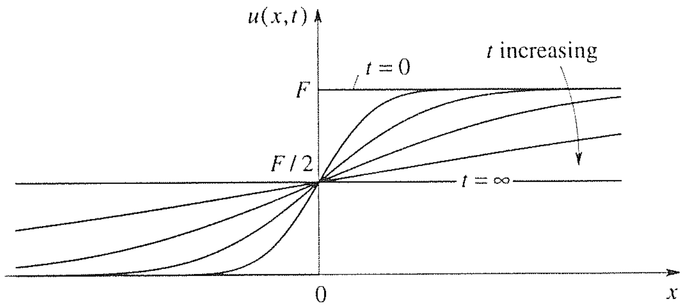


Figure 4. The solution (14).

COMMENT 3. It is instructive to write (11) in the form

$$u(x,t) = \int_{-\infty}^{\infty} f(\xi)K(\xi - x; t)\,d\xi, \tag{18a}$$

where*

$$K(\xi - x; t) = \frac{e^{-(x-\xi)^2/(4\alpha^2 t)}}{2\alpha\sqrt{\pi t}} \tag{18b}$$

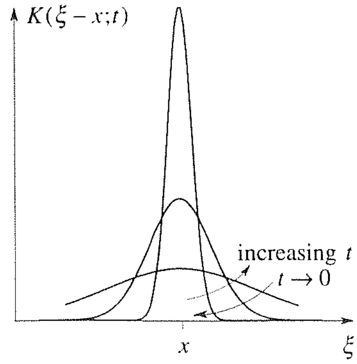


Figure 5. The delta-sequence behavior of K as $t \to 0$, and as t increases.

is called the **kernel**. That is, it comprises everything in the integrand other than the input $f(\xi)$. The kernel $K(\xi - x; t)$ happens to be a **normal** (or **Gaussian**) **probability distribution** centered at $\xi = x$, which result correctly suggests that diffusion is an essentially statistical phenomenon. The area between the graph of K and the ξ axis is unity for all t (Exercise 3), and that graph becomes more and more focused as $t \to 0$ (Fig. 5). If you studied Section 5.6 you will see that K appears to approach a delta function at x, $\delta(\xi - x)$, as $t \to 0$ because it has unit area for each t, and it becomes focused at x as $t \to 0$. In fact,

*Here, ξ is the active variable since it is the variable of integration; x and t are regarded as fixed in (18a). If we write $K(\xi; x, t)$ rather than $K(\xi, x, t)$, we are emphasizing that ξ is the active variable and that x and t are, at least for the moment, fixed. Even so, we have written $K(\xi - x; t)$ instead because only the difference $\xi - x$ occurs in (18b). We say that $K(\xi - x; t)$ is a **difference kernel**. The present example is similar to Example 8 of Section 17.10, which we urge you to review when you have finished reading this example.

that observation is confirmed by the initial condition, which is

$$\lim_{t\to 0} u(x,t) = \lim_{t\to 0}\int_{-\infty}^{\infty} f(\xi)K(\xi - x; t)\, d\xi = f(x). \tag{19}$$

That is, the initial condition (19) is satisfied *by virtue* of the fact that $K(\xi - x; t)$ tends to a delta function $\delta(\xi - x)$ as $t \to 0$.

COMMENT 4. Since the kernel K is evidently important, we would do well to try to understand its physical significance. To do so, let the input itself be a delta function at some point x_0,

$$f(x) = \delta(x - x_0). \tag{20}$$

In a crude way, this case can be conceptualized as corresponding to the application of a welding torch to the rod at x_0 for a brief instant. Then (18a) gives

$$\begin{aligned} u(x,t) &= \int_{-\infty}^{\infty} \delta(\xi - x_0)K(\xi - x; t)\, d\xi \\ &= K(x_0 - x; t) = K(x - x_0; t). \end{aligned} \tag{21}$$

The second equality in (21) follows from the fundamental property

$$\int_{-\infty}^{\infty} \delta(\xi - a)g(\xi)\, d\xi = g(a) \tag{22}$$

of the delta function, and the third equality in (21) follows from the fact that K is an even function of $x_0 - x$ [because it contains $(x_0 - x)^2$]. The upshot is that the kernel $K(x - x_0; t)$ is itself a solution of the heat conduction equation (1a), corresponding to an initial temperature distribution $\delta(x - x_0)$. Thus, $K(\xi - x; t)$, the graph of which is shown in Fig. 5, is the temperature distribution in the rod that results from an initial temperature distribution that is a delta function at x. Once again, we see the smoothing nature of the diffusion process, for beginning with the spike-like temperature profile $\delta(\xi - x)$ the temperature distribution $u(x,t)$ smooths out more and more as t increases (Fig. 5). Finally, we can now understand the superposition nature of (18a), for if $K(x - \xi; t)$ is the temperature response to an initial temperature that is a delta function (hence having unit area) at x, then the response $du(x,t)$ to the rectangular-pulse initial temperature shown in Fig. 6, having area $f(\xi)\, d\xi$, is $K(x - \xi; t)$ scaled by $f(\xi)\, d\xi$,

$$du(x,t) = K(x - \xi; t)[f(\xi)\, d\xi] = K(\xi - x; t)[f(\xi)\, d\xi]. \tag{23}$$

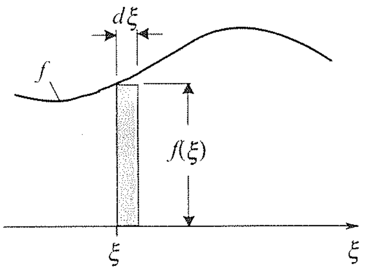


Figure 6. Breaking f into rectangular pulses.

Adding such results for all of the rectangular pulses that comprise f gives the integral (18a).

COMMENT 5. As a final observation about the physics, observe that the solution (18) indicates the spreading of information, by diffusion, with an *infinite velocity*. For instance, in Fig. 4 we see that some of the heat that was initially confined to the interval $0 < x < \infty$ diffuses to the interval $-\infty < x < 0$ over any arbitrarily small time t. This spreading can occur only if the speed of propagation is infinite. Inasmuch as it is agreed that energy cannot propagate faster than the speed of light, this result evidently reveals a flaw in our diffusion equation $\alpha^2 u_{xx} = u_t$, one that is discussed in the literature,[*] and which is inconsequential in practical applications of the theory.

COMMENT 6. We were able to use a Fourier transform on x to solve (1) because the latter was a boundary value problem on $-\infty < x < \infty$. Alternatively, we could have used a Laplace transform on t because (1) was an initial value problem on $0 < t < \infty$. The latter approach proves to be less attractive because we obtain a *non*homogeneous ODE on $\overline{u}(x, s)$ rather than the homogeneous ODE (3) on $\hat{u}(\omega, t)$. (See Exercise 5.) ▮

In Example 1 we illustrate the use of the Fourier transform for a problem on the conduction of heat in an infinite rod ($-\infty < x < \infty$). In Example 2 we illustrate the use of the Laplace transform for a semi-infinite rod ($0 < x < \infty$) problem.

EXAMPLE 2. *Heat Conduction in a Semi-Infinite Rod.* This time, we consider the problem

$$\alpha^2 u_{xx} = u_t, \qquad (0 < x < \infty,\ 0 < t < \infty) \tag{24a}$$

$$u(x,0) = 0, \qquad (0 < x < \infty) \tag{24b}$$

$$u(0,t) = g(t), \qquad (0 < t < \infty) \tag{24c}$$

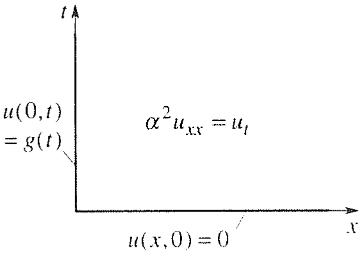


Figure 7. Semi-infinite rod problem.

as summarized in Fig. 7. The rod is initially at 0° throughout, we subject the left end ($x = 0$) to a prescribed temperature $g(t)$, and we seek the temperature distribution $u(x, t)$ that develops. We will also need a boundary condition at the right end ($x = \infty$), which we take to be $u(\infty, t) = 0$ for all t; that is,

$$\lim_{x\to\infty} u(x,t) = 0. \qquad (0 < t < \infty) \tag{24d}$$

In this case a Fourier transform is inappropriate because the domain is $0 < x < \infty$ rather than $-\infty < x < \infty$, so let us try a Laplace transform on t.[*] Thus, Laplace transform (24a) with respect to t:

$$L\{\alpha^2 u_{xx}\} = L\{u_t\}. \tag{25}$$

Now,

$$\begin{aligned} L\{\alpha^2 u_{xx}\} &= \alpha^2 L\{u_{xx}\} \qquad \text{(linearity of } L\text{)} \\ &= \alpha^2 \int_0^\infty \frac{\partial^2 u}{\partial x^2}\, e^{-st}\, dt \qquad \text{(definition of transform)} \\ &= \alpha^2 \frac{d^2}{dx^2} \int_0^\infty u(x,t) e^{-st}\, dt \qquad \text{(Leibniz rule)} \\ &= \alpha^2 \frac{d^2}{dx^2}\, \overline{u} \\ &= \alpha^2 \overline{u}_{xx}(x,s), \end{aligned} \tag{26}$$

[*] See, for example, P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Part I (New York: McGraw-Hill, 1953), pp. 865–869.

[*] Alternatively, we could use a semi-infinite Fourier transform, namely, a Fourier cosine or sine transform. These are studied in the optional Section 17.11.

and

$$L\{u_t\} = s\overline{u}(x,s) - u(x,0) \\ = s\overline{u}(x,s), \tag{27}$$

so (25) becomes $\alpha^2\overline{u}_{xx} = s\overline{u}$, or

$$\overline{u}_{xx} - \frac{s}{\alpha^2}\overline{u} = 0, \tag{28}$$

which is a second-order ODE with respect to x, t having been eliminated by the Laplace transform process. Solving (27),

$$\overline{u}(x,s) = Ae^{\sqrt{s}\,x/\alpha} + Be^{-\sqrt{s}\,x/\alpha}. \tag{29}$$

We have already used (24a) and (24b). To solve for A and B we use the boundary conditions (24c) and (24d), but first we need to express those conditions in terms of s rather than t. For the condition at $x = \infty$ observe that

$$\lim_{x\to\infty} \overline{u}(x,s) = \lim_{x\to\infty} \int_0^\infty u(x,t)e^{-st}\,dt \\ = \int_0^\infty \lim_{x\to\infty} u(x,t)e^{-st}\,dt = \int_0^\infty 0\,dt = 0; \tag{30}$$

that is, $u(x,t) \to 0$ as $x \to \infty$ implies that

$$\overline{u}(x,s) \to 0 \ \text{ as } \ x \to \infty \tag{31}$$

as well. [Note that the second equality in (30) was carried out only formally inasmuch as we did not rigorously justify the interchange in the order of the two limit processes: the limit $x \to \infty$ and the limit process that lies behind the Riemann integral.] Applying (31) to (29) gives $A = 0$, so

$$\overline{u}(x,s) = Be^{-\sqrt{s}\,x/\alpha}. \tag{32}$$

To express (24c) in terms of s, take its Laplace transform:

$$\int_0^\infty u(0,t)e^{-st}\,dt = \int_0^\infty g(t)e^{-st}\,dt,$$

or

$$\overline{u}(0,s) = \overline{g}(s), \tag{33}$$

and imposing the latter upon (32) gives

$$\overline{u}(0,s) = \overline{g}(s) = Be^0, \tag{34}$$

so $B = \overline{g}(s)$ and

$$\overline{u}(x,s) = \overline{g}(s)e^{-\sqrt{s}\,x/\alpha}. \tag{35}$$

To invert, use entry 21 of Appendix C [with $a = x/(2\alpha)$] and the convolution property (entry 28):

$$u(x,t) = g(t) * \frac{xe^{-x^2/(4\alpha^2 t)}}{2\alpha\sqrt{\pi}\,t^{3/2}} \\ = \frac{x}{2\alpha\sqrt{\pi}} \int_0^t g(t-\tau)\,\frac{e^{-x^2/(4\alpha^2\tau)}}{\tau^{3/2}}\,d\tau, \tag{36}$$

or equivalently,

$$u(x,t) = \frac{x}{2\alpha\sqrt{\pi}} \int_0^t g(\tau)\, \frac{e^{-x^2/[4\alpha^2(t-\tau)]}}{(t-\tau)^{3/2}}\, d\tau \tag{37}$$

if we prefer.

COMMENT 1. The application of (31) to the solution (29) was easy: the positive exponential grows with x and the negative exponential dies out, so $A = 0$ and B remained arbitrary. What if we had used the solution form

$$\overline{u}(x,s) = C \cosh \frac{\sqrt{s}\, x}{\alpha} + D \sinh \frac{\sqrt{s}\, x}{\alpha} \tag{38}$$

instead of (29), for *both* terms on the right-hand side tend to infinity as $x \to \infty$? Let us see:

$$\begin{aligned}\overline{u}(x,s) &= \frac{C}{2}\left(e^{\sqrt{s}\,x/\alpha} + e^{-\sqrt{s}\,x/\alpha}\right) + \frac{D}{2}\left(e^{\sqrt{s}\,x/\alpha} - e^{-\sqrt{s}\,x/\alpha}\right) \\ &= \frac{C+D}{2} e^{\sqrt{s}\,x/\alpha} + \frac{C-D}{2} e^{-\sqrt{s}\,x/\alpha}.\end{aligned} \tag{39}$$

To eliminate the positive exponential, set $D = -C$. Then (39) becomes $\overline{u}(x,s) = C \exp -(\sqrt{s}\, x/\alpha)$, which is equivalent to (32). The point is that either form will work, (29) or (38), but (29) is more convenient insofar as the application of the condition (31).

COMMENT 2. We did not specify the function $g(t)$. Observe that if $g(t)$ *were* specified, it would have been foolish to work out its transform $\overline{g}(s)$ because when we apply the convolution theorem to (35) we invert $\overline{g}(s)$ and get back the given function $g(t)$. Thus, it is best to merely call the transform $\overline{g}(s)$, as we did.

COMMENT 3. Observe from (37) that $u(x,t)$ depends upon the boundary data $g(t)$ only over $0 < \tau < t$, not over $0 < \tau < \infty$. This result is entirely reasonable since how could the temperature distribution $u(x,t)$ today depend upon the boundary values $g(t)$ to be imposed tomorrow?

COMMENT 4. Finally, let us use (37) to determine $u(x,t)$ for a specific case, say $g(t) = 100°$. With $g(t)$ a constant, it is easier to use (36) than (37), and the latter gives (Exercise 13)

$$\begin{aligned}u(x,t) &= \frac{100x}{2\alpha\sqrt{\pi}} \int_0^t \frac{e^{-x^2/(4\alpha^2\tau)}}{\tau^{3/2}}\, d\tau \\ &= 100\, \text{erfc}\left(\frac{x}{2\alpha\sqrt{t}}\right),\end{aligned} \tag{40}$$

which is plotted in Fig. 8 at representative times. ▮

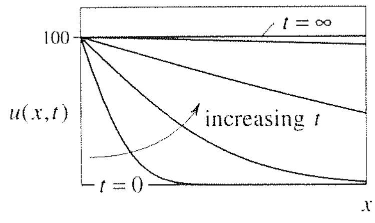


Figure 8. $u(x,t)$ given by (40).

Closure. In Section 18.3 we learn how to solve the diffusion equation by separation of variables, but the x domain is always finite and the boundary conditions do not vary with t (although certain more complicated cases are outlined in the end-of-section exercises). The Fourier and Laplace transforms enable us to deal with semi-infinite ($0 < x < \infty$) and infinite ($-\infty < x < \infty$) domains, as well as nonconstant boundary conditions.

Application of the Laplace and Fourier transform to the diffusion equation $\alpha^2 u_{xx} = u_t$ proceeds very much along the same lines as for ordinary differential equations, but instead of producing an algebraic equation it produces an ordinary differential equation. Specifically, Laplace transforming on t produces the ODE $\alpha^2 \overline{u}_{xx} - s\overline{u} = -u(x,0)$ on $\overline{u}(x,s)$, and Fourier transforming on x (if $-\infty < x < \infty$) produces the ODE $\hat{u}_t + \alpha^2\omega^2\hat{u} = 0$ on $\hat{u}(\omega, t)$.

EXERCISES 18.4

NOTE: Exercises 1–10 relate to Example 1; the remainder to Example 2.

1. Show that for the case where $f(x) = \text{constant} \equiv F$, (11) gives $u(x,t) = F$. HINT: Make the change of variables $(x-\xi)^2/(4\alpha^2 t) = \mu^2$ in the integral, and use the known integral $\int_{-\infty}^{\infty} \exp{(-\xi^2)}\, d\xi = \sqrt{\pi}$.

2. (a) Show that for the case where $f(x) = FH(x)$, in (13), (11) gives (14) as the solution.
(b) Verify, directly, that (14) satisfies the PDE (1a) and the initial condition (1b).

3. Prove the claim, made in Comment 3 of Example 1, that $\int_{-\infty}^{\infty} K(\xi - x; t)\, d\xi = 1$ for all t, where K is given by (18b). HINT: Use the change of variables suggested in Exercise 1.

4. Show that $\operatorname{erf}(x)$, defined by (15), is an odd function, as is claimed in Comment 2 of Example 1.

5. We used the Fourier transform to solve (1). Use the Laplace transform instead, and obtain the ODE

$$\overline{u}_{xx} - \frac{s}{\alpha^2}\overline{u} = -\frac{1}{\alpha^2} f(x).$$

6. Verify, by direct substitution, that the kernel K given by (18b) satisfies the diffusion equation (1a), as was claimed in Example 1.

7. Use (18) to show that if the initial temperature $f(x)$ is
(a) a periodic function of x, with period T, then so is the solution $u(x,t)$.
(b) an odd function of x, then so is the solution $u(x,t)$.
(c) an even function of x, then so is the solution $u(x,t)$.

8. (*Inclusion of a source term*) In Example 3 of Section 16.8 we find that if there is a heat source distribution $F(x,t)$ within the medium, then in place of the homogeneous field equation $L[u] = \alpha^2 u_{xx} - u_t = 0$ we have the nonhomogeneous equation $L[u] = \alpha^2 u_{xx} - u_t = -F(x,t)$; F acts as a source where $F > 0$ and as a sink where $F < 0$.

(a) Then the problem

$$\begin{array}{rll} \alpha^2 u_{xx} - u_t &= -F(x,t), & (-\infty < x < \infty,\ \ 0 < t < \infty) \\ u(x,0) &= f(x), & (-\infty < x < \infty) \end{array} \tag{8.1}$$

together with suitable boundary conditions at $x = \pm\infty$, has two inputs, the initial temperature $f(x)$ and the source distribution $F(x,t)$. By linearity, the response $u(x,t)$ should be the sum of the responses to these individual inputs. Specifically, show that if $v(x,t)$ and $w(x,t)$ are solutions to the problems

$$\begin{array}{rll} \alpha^2 v_{xx} - v_t &= 0, & (-\infty < x < \infty,\ \ 0 < t < \infty) \\ v(x,0) &= f(x) & (-\infty < x < \infty) \end{array} \tag{8.2}$$

and

$$\begin{array}{rll} \alpha^2 w_{xx} - w_t &= -F(x,t), & (-\infty < x < \infty,\ \ 0 < t < \infty) \\ w(x,0) &= 0, & (-\infty < x < \infty) \end{array} \tag{8.3}$$

respectively, then $u(x,t) = v(x,t) + w(x,t)$. Since the v problem is already solved in Example 1, the remainder of this exercise is devoted to the w problem.
(b) Consider the case where $F = F(t)$ is a function of t alone. Choosing between the Fourier and Laplace transforms, solve for w. (The answer should be in the form of an integral.) Explain why you selected the transform that you did, and not the other.
(c) Now consider the case where $F = F(x)$ is a function of x alone. Using a Fourier transform, show that

$$w(x,t) = F(x) * F^{-1}\left\{\frac{1-e^{-\alpha^2\omega^2 t}}{\alpha^2\omega^2}\right\}. \tag{8.4}$$

(d) The inverse needed in (8.4) is not found in our brief table (Appendix D). Nonetheless, show that

$$F^{-1}\left\{\frac{1-e^{-\alpha^2\omega^2 t}}{\alpha^2\omega^2}\right\} = \frac{1}{\alpha}\sqrt{\frac{t}{\pi}}\,e^{-x^2/(4\alpha^2 t)} - \frac{x}{2\alpha^2}\operatorname{erfc}\left(\frac{x}{2\alpha\sqrt{t}}\right), \tag{8.5}$$

so

$$w(x,t) = \frac{1}{\alpha}\int_{-\infty}^{\infty} F(x-\xi)\left[\sqrt{\frac{t}{\pi}}\,e^{-\xi^2/(4\alpha^2 t)} - \frac{\xi}{2\alpha}\operatorname{erfc}\left(\frac{\xi}{2\alpha\sqrt{t}}\right)\right] d\xi. \tag{8.6}$$

HINT: Letting $\hat{g} = (1-e^{-\alpha^2\omega^2 t})/(\alpha^2\omega^2)$, show that $\hat{g}_t = e^{-\alpha^2\omega^2 t} = \hat{g}_t$, so that $g_t = \{\exp[-x^2/(4\alpha^2 t)]\}/(2\alpha\sqrt{\pi t})$, with $g|_{t=0} = 0$. Thus,

$$g(x,t) = \int_0^t \frac{e^{-x^2/(4\alpha^2\tau)}}{2\alpha\sqrt{\pi\tau}}\,d\tau.$$

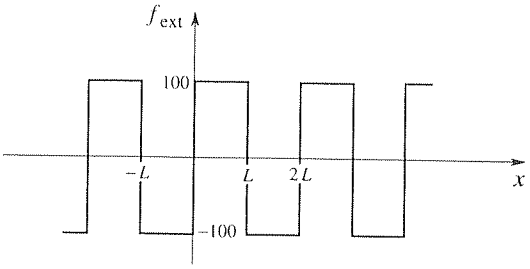
Then, use change of variables and integration by parts.

9. (*Small-t solution for finite rod*) In Section 18.3 we obtain, by separation of variables, the solution

$$u(x,t) = \frac{400}{\pi}\sum_{n=1,3,\dots}^{\infty}\frac{1}{n}\sin\frac{n\pi x}{L}\,e^{-(n\pi\alpha/L)^2 t} \tag{9.1}$$

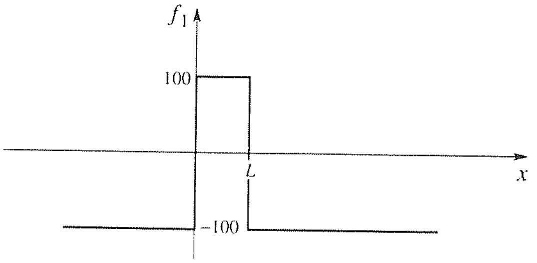
to the problem of heat conduction in a rod of length L initially at a uniform temperature $u(x,0) = 100°$, with both ends held subsequently at $0°$. It was pointed out that (9.1) converges rapidly if t is large and slowly if t is small. Our object in this exercise is to show how to use the results obtained in this section to obtain a complementary result: a series solution that *converges rapidly for small* t (and slowly for large t). First, observe that our use of the half-range sine series to solve the stated finite-rod problem, on $0 < x < L$, is equivalent to solving the periodically extended *infinite*-rod problem

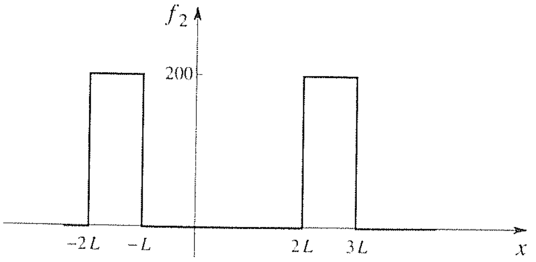
$$\alpha^2 u_{xx} = u_t, \qquad (-\infty < x < \infty,\ 0 < t < \infty)$$
$$u(x,0) = f_{\text{ext}}(x),$$



where f_{ext} is the square wave shown here. In effect, the separation-of-variables solution amounts to expanding f_{ext} in a Fourier sine series, finding the response due to each sine term, and then adding them. Alternatively, suppose that we expand $f_{\text{ext}}(x)$ as $f_{\text{ext}}(x) = f_1(x) + f_2(x) + \cdots$, where the f_j's are as shown. That is,

$$\begin{aligned} f_{\text{ext}}(x) = {} & \{-100 + 200[H(x) - H(x-L)]\} \\ & +200\{H(x+2L) - H(x+L) \\ & +H(x-2L) - H(x-3L)\} \\ & +200\{H(x+4L) - H(x+3L) \\ & +H(x-4L) - H(x-5L)\} + \cdots. \end{aligned} \tag{9.2}$$





(a) Recalling the solution (14), for the case where $f(x)$ is given by (13), show that the responses to $f_1, f_2, \dots$ are

$$u_1(x,t) = 100\left[\operatorname{erf}\left(\frac{x}{2\alpha\sqrt{t}}\right) - \operatorname{erf}\left(\frac{x-L}{2\alpha\sqrt{t}}\right) - 1\right] \tag{9.3}$$

and

$$u_j(x,t) = 100\left[\operatorname{erf}\left(\frac{x+2(j-1)L}{2\alpha\sqrt{t}}\right) - \operatorname{erf}\left(\frac{x+(2j-3)L}{2\alpha\sqrt{t}}\right) + \operatorname{erf}\left(\frac{x-2(j-1)L}{2\alpha\sqrt{t}}\right) - \operatorname{erf}\left(\frac{x-(2j-1)L}{2\alpha\sqrt{t}}\right)\right], \tag{9.4}$$

for $j \geq 2$.

(b) Explain, in simple physical terms, why the series solution

$$u(x,t) = u_1(x,t) + u_2(x,t) + \cdots \tag{9.5}$$

should converge rapidly for small t.

(c) To illustrate, compute $u(x,t)$ at $x = 1$, $t = 0.1$, with $\alpha^2 = 1.14$ and $L = 10$, using equation (9.1), and again, using equation (9.5). Two-significant-figure accuracy will suffice. NOTICE: To calculate the error functions, use either the Hastings formula that we gave in a footnote, or use computer software such as the evalf *Maple* command.

(d) In (b) we stated that (9.5) should converge rapidly for small t. What do we really mean by small t? $t < 1$ sec? $t < [L/(4\alpha)]^2$? Propose some such inequality that can be used as a guarantee that (9.5) will indeed be rapidly convergent.

(e) Obtain a computer plot of the solution at $t = 0.1, 0.5$, and 1, using the approximation $u(x,t) \approx u_1(x,t)$, where u_1 is given in (9.3). Take $\alpha^2 = 1.14$ and $L = 10$, as in part (c).

10. (*Translating rod*) We saw in Section 18.2 that if the rod is translating rightward with constant speed v, then the PDE becomes $\alpha^2 u_{xx} = u_t + V u_x$, where $V = v/c$ and c is the specific heat of the material. Use the Fourier transform to solve the problem.

$$\begin{aligned} \alpha^2 u_{xx} &= u_t + V u_x, && (-\infty < x < \infty,\ 0 < t < \infty) \\ u(x,0) &= f(x), && (-\infty < x < \infty) \end{aligned}$$

where $u \to 0$ and $u_x \to 0$ as $x \to \pm\infty$. Of course, your result should reduce to (18) for the case where $V = 0$.

11. Rework Example 2 with the initial condition changed to $u(x,0) = \text{constant} = u_0$, and show that

$$u(x,t) = u_0 + (u_1 - u_0)\operatorname{erfc}\left(\frac{x}{2\alpha\sqrt{t}}\right).$$

12. (*Oscillatory temperature at the left end*) If an oscillatory temperature is maintained at the left end of a semi-infinite rod, then we expect the solution $u(x,t)$ to approach a steady-state oscillation. Specifically, we have

$$\begin{aligned} \alpha^2 u_{xx} &= u_t, && (0 < x < \infty,\ 0 < t < \infty) \\ u(0,t) &= u_0 \cos \omega t, && (0 < t < \infty) \\ u &\to 0 \text{ as } x \to \infty. && (0 < t < \infty) \end{aligned} \tag{12.1}$$

Derive the steady-state oscillatory solution

$$u(x,t) = u_0 e^{-rx} \cos(\omega t - rx), \tag{12.2}$$

where $r = \alpha\sqrt{\omega/2}$. HINT: The simplest approach is to consider, instead, the problem

$$\begin{aligned} \alpha^2 v_{xx} &= v_t && (0 < x < \infty,\ 0 < t < \infty) \\ v(0,t) &= u_0 e^{i\omega t}, && (0 < t < \infty) \\ v &\to 0 \text{ as } x \to \infty, && (0 < t < \infty) \end{aligned} \tag{12.3}$$

which can then be solved by seeking v in the form

$$v(x,t) = X(x)e^{i\omega t}. \tag{12.4}$$

Then u is found as the real part of v:

$$u(x,t) = \operatorname{Re} v(x,t). \tag{12.5}$$

Such a **complex function method** of solution, for differential equations with oscillatory forcing terms, is the subject of Exercise 12 in Section 3.8. Alternatively, we could anticipate the phase shift caused by the u_t term in the PDE and seek u, directly, in the form

$$u(x,t) = A(x)\cos \omega t + B(x)\sin \omega t,$$

but the complex function method is more attractive since it permits us to work with a single quantity $e^{i\omega t}$ rather than with a cosine and sine. Observe that there is no initial condition $u(x,0)$ included in (12.1) because we are concerned here only with the steady-state solution. NOTE: Besides the heat conduction application considered here, the problem (12.1) also arises (with the t's changed to y's) in fluid mechanics regarding the viscous flow, in the upper half plane $y > 0$, that is caused by harmonic oscillation of an infinite flat plate at $y = 0$. There, the problem is known as Stokes's second problem and was studied also by Lord Rayleigh.

13. (a) Show that the integral in (40) does give $u(x,t) = 100 \operatorname{erfc}(x/(2\alpha\sqrt{t}))$.

(b) Use computer software to evaluate the latter at $x = 0, 2, 4, 6, \ldots, 20$, with $t = 10$ and $\alpha^2 = 1.14$.

14. (*Heat flux at left end*) The problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t, \qquad (0 < x < \infty,\ 0 < t < \infty) \\ u(x,0) &= 0, \qquad (0 < x < \infty) \\ u_x(0,t) &= -Q, \qquad (0 < t < \infty) \end{aligned} \tag{14.1}$$

with $u \to 0$ as $x \to \infty$, differs from (24) in that the boundary condition at $x = 0$ is of Neumann type rather than of Dirichlet type. Physically, $u_x(0,t) = -Q$ (where Q is a prescribed constant) corresponds to the maintaining of a constant heat flux into the rod at $x = 0$, for all $t > 0$ (i.e., into the rod if $Q > 0$, out of the rod if $Q < 0$).

(a) Using the Laplace transform, derive the result

$$u(x,t) = \frac{\alpha Q}{\sqrt{\pi}} \int_0^t \frac{e^{-x^2/(4\alpha^2\tau)}}{\sqrt{\tau}}\, d\tau. \tag{14.2}$$

In particular, use (14.2) to solve for the temperature at the left end, $u(0,t)$, and sketch its graph.

(b) Show that the integral in (14.2) can be evaluated, and that we obtain

$$u(x,t) = Q\left[2\alpha\sqrt{\frac{t}{\pi}}e^{-x^2/(4\alpha^2 t)} - x\,\mathrm{erfc}\left(\frac{x}{2\alpha\sqrt{t}}\right)\right]. \tag{14.3}$$

15. We claimed, in a footnote, that Hastings's approximate formula for $\mathrm{erf}(x)$ is uniformly accurate, over $0 \le x < \infty$, to ± 0.000025. Check that claim for $x = 0.5, 1,$ and 2 by using that formula to compute $\mathrm{erf}(x)$, and comparing those results with results obtained either from computer software (such as the *Maple* evalf command) or from tables.

18.5 The Method of Images (Optional)

The method of images is a method of fictitiously extending the problem domain so as to satisfy homogeneous boundary conditions by means of symmetries or antisymmetries. In this section we not only illustrate the method for the diffusion equation, we also establish a class of PDE's for which the method works.

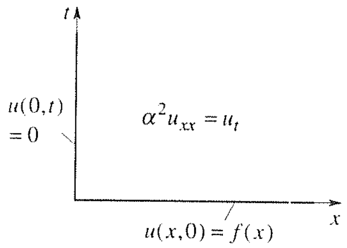


Figure 1. The problem (1).

18.5.1. Illustration of the method. To illustrate the method of images, consider the diffusion problem

$$\alpha^2 u_{xx} = u_t, \qquad (0 < x < \infty,\ 0 < t < \infty) \tag{1a}$$
$$u(x,0) = f(x), \qquad (0 < x < \infty) \tag{1b}$$
$$u(0,t) = 0, \qquad (0 < t < \infty) \tag{1c}$$

where $u \to 0$ and $u_x \to 0$ as $x \to \infty$ as depicted in Fig. 1.

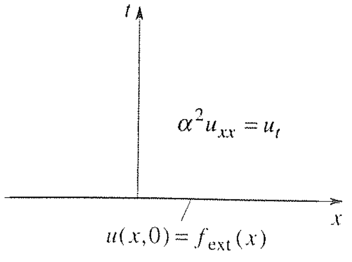


Figure 2. Extended problem.

The idea behind the image method is to extend the problem domain to $x = -\infty$ as shown in Fig. 2. In doing so, we also need to extend the initial condition to $x = -\infty$. Calling the extended function f_{ext}, we need $f_{\text{ext}}(x) = f(x)$ for $x > 0$, but for $x < 0$ we can define $f_{\text{ext}}(x)$ in any way we choose. Let us choose $f_{\text{ext}}(x)$ to be the *odd* extension of $f(x)$. For instance, if $f(x)$ is as shown in Fig. 3a, then $f_{\text{ext}}(x)$ is as shown in Fig. 3b. With $u(x,0) = f_{\text{ext}}(x)$ being an odd function of x (i.e., antisymmetric about $x = 0$), we expect $u(x,t)$ to *remain* an odd function of x for all $t > 0$, * and if $u(x,t)$ is an odd function of x for all $t > 0$ then $u(0,t) = 0$ for all $t > 0$. That is, by building antisymmetry about $x = 0$ into the extended

*If $u(x,t)$ is an odd function of x, then $u(-x,t) = -u(x,t)$, so $u(0,t) = -u(0,t)$. Hence, $2u(0,t) = 0$, so $u(0,t) = 0$.

problem we are able to automatically satisfy the $u(0,t) = 0$ boundary condition in (1c).

Thus, the solution of the extended problem (Fig. 2) is also the solution of the original problem (Fig. 1). In fact, the extended problem was already solved in Example 1 of Section 18.4, so the desired solution of (1) is given by

$$u(x,t) = \int_{-\infty}^{\infty} f_{\text{ext}}(\xi) K(\xi - x; t)\, d\xi, \tag{2a}$$

where

$$K(\xi - x; t) = \frac{e^{-(\xi - x)^2/(4\alpha^2 t)}}{2\alpha\sqrt{\pi t}}. \tag{2b}$$

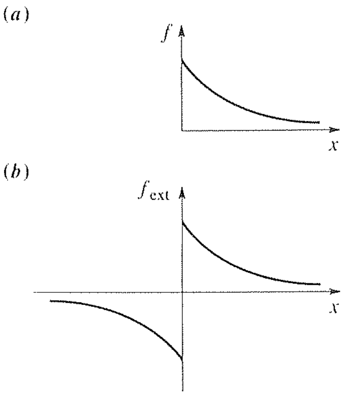


Figure 3. Odd extension.

Since we use (2) to compute u only over the actual domain, $0 < x < \infty$, it is preferable (though not necessary) to re-express (2) so as to eliminate all reference to the fictitious extension, which extension is referred to as the **image system**. To do so, write

$$\begin{aligned}
u(x,t) &= \int_{-\infty}^{0} f_{\text{ext}}(\xi) K(\xi - x; t)\, d\xi + \int_{0}^{\infty} f_{\text{ext}}(\xi) K(\xi - x; t)\, d\xi \\
&= \int_{\infty}^{0} f_{\text{ext}}(-\mu) K(-\mu - x; t)\, (-d\mu) + \int_{0}^{\infty} f_{\text{ext}}(\xi) K(\xi - x; t)\, d\xi \\
&= -\int_{\infty}^{0} [-f_{\text{ext}}(\mu)] K(\mu + x; t)\, d\mu + \int_{0}^{\infty} f_{\text{ext}}(\xi) K(\xi - x; t)\, d\xi, \qquad (3)
\end{aligned}$$

where we set $\xi = -\mu$ in the first integral and used the oddness of f_{ext} [i.e., $f_{\text{ext}}(-\mu) = -f_{\text{ext}}(\mu)$] and the evenness of K [i.e., $K(-\mu - x; t) = K(\mu + x; t)$] in the fifth integral. Finally, set $\mu = \xi$ and note that $f_{\text{ext}}(\xi) = f(\xi)$ over $0 < \xi < \infty$, and obtain

$$\begin{aligned}
u(x,t) &= \int_{0}^{\infty} f(\xi)[K(\xi - x; t) - K(\xi + x; t)]\, d\xi \\
&= \int_{0}^{\infty} f(\xi) \frac{e^{-\frac{(\xi - x)^2}{4\alpha^2 t}} - e^{-\frac{(\xi + x)^2}{4\alpha^2 t}}}{2\alpha\sqrt{\pi t}}\, d\xi, \qquad (4)
\end{aligned}$$

which, as we desired, contains no reference to the image system.

Recall that we said that we "expect" $u(x,t)$ to remain an odd function of x for all $t > 0$ and to thereby satisfy the condition (1c), that $u(0,t) = 0$ for all $t > 0$. With (4) in hand, we can now verify that claim since (4) gives

$$u(0,t) = \int_{0}^{\infty} f(\xi)(0)\, d\xi = 0. \tag{5}$$

EXAMPLE 1. For instance, let $f(x) = 100$ in (4). Then (4) gives (Exercise 1)

$$u(x,t) = 100\,\text{erf}\left(\frac{x}{2\alpha\sqrt{t}}\right). \tag{6}$$

We've plotted (6), at representative times, in Fig. 4, along with the fictitious extension (shown as dashed). Observe how the antisymmetry of u, with respect to x, ensures the satisfaction of the boundary condition $u(0,t) = 0$ for all $t > 0$. ∎

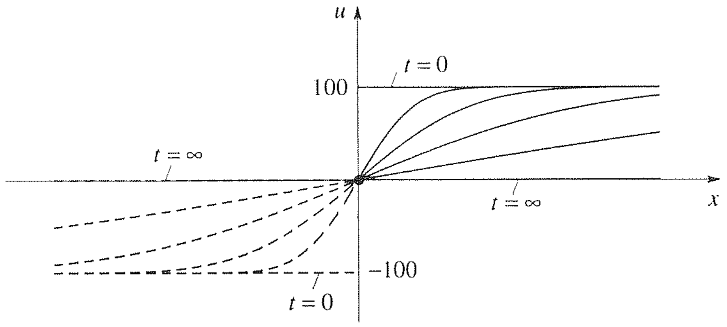


Figure 4. Solution for $f(x) = 100$.

Suppose that instead of the homogeneous Dirichlet condition (1c) we have the homogeneous Neumann condition

$$u_x(0,t) = 0. \qquad (0 < t < \infty) \tag{7}$$

In physical terms, instead of applying ice to the left end of the rod we insulate it so there is no heat flux across the face $x = 0$. In that case we extend $f(x)$ so as to be an *even* function, symmetric about $x = 0$. Then, we expect $u(x,t)$ to be a symmetric function of x for all $t > 0$ so that, by virtue of that symmetry, we will have $u_x(0,t) = 0$ for all $t > 0$ (Exercise 2).

18.5.2. Mathematical basis for the method. The key point, in applying the method of images is that if

$$L[u] = \alpha^2 u_{xx} - u_t = 0 \qquad (-\infty < x < \infty, \; 0 < t < \infty) \tag{8a}$$

and

$$u(x,0) = f(x), \qquad (-\infty < x < \infty) \tag{8b}$$

together with suitable boundary conditions at $x = \pm\infty$, then $f(x)$ being odd implies that the solution $u(x,t)$ is an odd function of x for all $t > 0$, and $f(x)$ being even implies that $u(x,t)$ is an even function of x for all $t > 0$. Let us explain the mathematical basis for that claim, not just for the diffusion equation but for other PDE's as well.

We will draw upon the following elementary results, proof of which are left for the exercises.

1. Any function $F(x)$ can be split into even and odd parts, $F_e(x)$ and $F_o(x)$, respectively, as

$$F(x) = \underbrace{\frac{F(x) + F(-x)}{2}}_{F_e(x)} + \underbrace{\frac{F(x) - F(-x)}{2}}_{F_o(x)}. \tag{9}$$

2. The algebra of even and odd functions is as follows:

$$\begin{aligned} \text{even} \times \text{even} &= \text{even}, \\ \text{even} \times \text{odd} &= \text{odd}, \\ \text{odd} \times \text{odd} &= \text{even}. \end{aligned} \tag{10}$$

3. If $E_1(x)$ and $E_2(x)$ are even, and $O_1(x)$ and $O_2(x)$ are odd, then

$$E_1(x) + O_1(x) = E_2(x) + O_2(x) \tag{11}$$

implies that

$$E_1(x) = E_2(x) \qquad \text{and} \qquad O_1(x) = O_2(x). \tag{12}$$

4. If $F(x)$ is even, then $F'(0) = 0$.
5. If $F(x)$ is odd, then $F(0) = 0$.
6. If $F(x)$ is even, then $F(x), F''(x), F''''(x), \ldots$ are even and $F'(x)$, $F'''(x), \ldots$ are odd.
7. If $F(x)$ is odd, then $F(x), F''(x), F''''(x), \ldots$ are odd and $F'(x)$, $F'''(x), \ldots$ are even.

The foregoing results hold even if F depends on other variables as well, such as t. For instance, if $F(x,t)$ is an even function of x, then $F(x,t), F_{xx}(x,t), F_{xxxx}(x,t)$, $\ldots$ are even functions of x, and $F_x(x,t), F_{xxx}(x,t), \ldots$ are odd functions of x.

To proceed, it will be useful to consider the problem

$$L[u] = \alpha^2 u_{xx} - V u_x - u_t = Q(x,t), \qquad (|x| < \infty,\ t > 0) \tag{13a}$$

$$u(x,0) = f(x), \qquad (|x| < \infty) \tag{13b}$$

which is more general than (1) by virtue of including the Vu_x term (associated with translation of the rod at a constant speed) and the Q term (associated with the presence of a distributed heat source or sink along the rod). Of course, (13) reduces to (1) if we set $V = 0$ and $Q(x,t) = 0$.

To track the development of the even and odd parts of u, let us break u into the sum of its even and odd parts, say E and O, respectively:

$$u(x,t) = E(x,t) + O(x,t). \tag{14}$$

Putting (14) into (13a), and also splitting $Q(x,t)$ and $f(x)$ into their even and odd parts, gives

$$\underbrace{\alpha^2 E_{xx}}_{e} + \underbrace{\alpha^2 O_{xx}}_{o} - \underbrace{VE_x}_{o} - \underbrace{VO_x}_{e} - \underbrace{E_t}_{e} - \underbrace{O_t}_{o} = \underbrace{Q_e}_{e} + \underbrace{Q_o}_{o}, \tag{15}$$

where e or o shown below each term indicates whether that term is even or odd, respectively. For instance, the $\alpha^2 O_{xx}$ term is odd according to item 7 above, and the $-VE_x$ term is odd according to item 6. Partial differentiation with respect to t does not alter evenness or oddness (Exercise 7), so E_t is even and O_t is odd. Next, from item 3 it follows from (15) that $\alpha^2 E_{xx} - VO_x - E_t = Q_e$, and $\alpha^2 O_{xx} - VE_x - O_t = Q_o$. Similarly, it follows from (13b) that $E(x, 0) = f_e(x)$ and $O(x, 0) = f_o(x)$, so (13) can be split into the two problems

$$\alpha^2 E_{xx} - E_t = Q_e + VO_x, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \tag{16a}$$
$$E(x, 0) = f_e(x), \qquad (-\infty < x < \infty) \tag{16b}$$

and

$$\alpha^2 O_{xx} - O_t = Q_o + VE_x, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \tag{17a}$$
$$O(x, 0) = f_o(x). \qquad (-\infty < x < \infty) \tag{17b}$$

Suppose, first, that $V = 0$. Then (16) and (17) are uncoupled: (16) contains only $E(x, t)$ and (17) contains only $O(x, t)$. If both inputs f and Q are even, then $f_o = 0$ and $Q_o = 0$, the problem (17) on O is homogeneous, and $O(x, t) \equiv 0$. *Thus, if the inputs are even, then the solution $u(x, t)$ is even.* If, on the other hand, f and Q are odd, then $f_e = 0$ and $Q_e = 0$, the problem on E is homogeneous, and $E(x, t) \equiv 0$. *Thus, if the inputs are odd, then the solution $u(x, t)$ is odd.* The two italicized results enable us to use the method of images, as we did in Section 18.5.1.

However, if $V \neq 0$ then (16) and (17) are coupled by virtue of the VO_x and VE_x terms. For instance, even if f and Q are even, so $f_o = 0$ and $Q_o = 0$, (17) is still nonhomogeneous due to the VE_x, which acts as a source term and causes the development of a nonzero $O(x, t)$. Thus, we cannot use the method of images if $V \neq 0$, which result makes sense physically since translation of the rod (to the right if $V > 0$ and to the left if $V < 0$) will surely destroy any symmetry or antisymmetry in the solution about $x = 0$.

The upshot is that sometimes we can use the method of images and sometimes we cannot, depending on the operator L. What we need for the method of images to work is for L to be both *linear* and *even*. We say that an operator L is **even** if $L[u]$ is even whenever u is even and odd whenever u is odd; that is, an even operator preserves evenness and oddness.

Let us limit our attention to the linear operator

$$L = A\frac{\partial^2}{\partial x^2} + B\frac{\partial^2}{\partial t^2} + C\frac{\partial^2}{\partial x \partial t} + D\frac{\partial}{\partial x} + E\frac{\partial}{\partial t} + F, \tag{18}$$

where $A, \dots, F$ are functions of x and t. From the discussion above, we can see that L will be an even operator if A, B, E, F are even functions of x and if C, D are odd functions of x. In that case, let us re-express L as

$$\boxed{L = E_1\frac{\partial^2}{\partial x^2} + E_2\frac{\partial^2}{\partial t^2} + O_1\frac{\partial^2}{\partial x \partial t} + O_2\frac{\partial}{\partial x} + E_3\frac{\partial}{\partial t} + E_4.} \tag{19}$$

This case is typical rather than exceptional.

EXAMPLE 2. For instance, consider the classical one-dimensional wave equation $c^2u_{xx} = u_{tt}$ (studied in Chapter 19). In this case

$$L = c^2\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial t^2}, \tag{20}$$

which is even because $A = c^2$ and $B = -1$ are even, $C = D = 0$ are odd, and $E = F = 0$ are even. (Remember that the zero function is both even and odd; it is even because its odd part is zero, and it is odd because its even part is zero.) ▮

THEOREM 18.5.1 *Applicability of the Method of Images*
Let L be of the form (19) and therefore an even operator. Suppose that the problem

$$L[u] = Q(x,t), \qquad (-\infty < x < \infty, \;\; 0 < t < \infty) \tag{21a}$$
$$u(x,0) = f(x), \qquad (-\infty < x < \infty) \tag{21b}$$

where $u \to 0$ and $u_x \to 0$ as $x \to \pm\infty$, admits a unique solution $u(x,t)$. If f and Q are even functions of x then so is u, and if f and Q are odd functions of x then so is u.

Proof: Let $u(x,t) = E(x,t) + O(x,t)$, where E is even and O is odd. Then, $L[u] = L[E + O] = L[E] + L[O]$ because L is linear. Further, $L[E]$ is even and $L[O]$ is odd because L is even. Thus, (21) splits into the problems

$$L[E] = Q_e(x,t), \qquad (-\infty < x < \infty, \;\; 0 < t < \infty) \tag{22a}$$
$$E(x,0) = f_e(x), \qquad (-\infty < x < \infty) \tag{22b}$$

and

$$L[O] = Q_o(x,t), \qquad (-\infty < x < \infty, \;\; 0 < t < \infty) \tag{23a}$$
$$O(x,0) = f_o(x), \qquad (-\infty < x < \infty) \tag{23b}$$

where E, E_x, O, and O_x tend to zero as $x \to \pm\infty$. If f and Q are even, then $f_o = 0$ and $Q_o = 0$, and (23) gives the unique solution $O(x,t) = 0$, so $u(x,t)$ is even. If f and Q are odd, then $f_e = 0$ and $O_e = 0$ and (22) gives the unique solution $E(x,t) = 0$, so $u(x,t)$ is odd. ■

Closure. The method of images involves the fictitious extension of the problem domain such that homogeneous boundary conditions, which fall within the interior of the extended domain, are automatically satisfied by virtue of the symmetry or antisymmetry of the various inputs. For the method to work, we need the symmetry (or antisymmetry) of the inputs to imply the symmetry (or antisymmetry) of the output u. Such symmetry (or antisymmetry) will, indeed, be passed on to u if the operator L is linear and even where, by L being even, we mean that if u is even then $L[u]$ is even and if u is odd then $L[u]$ is odd. The linear operator L in (18) is even if it is of the form (19), where $E_1(x,t), \dots, E_4(x,t)$ are even functions of x and $O_1(x,t), O_2(x,t)$ are odd functions of x.

EXERCISES 18.5

1. Show that if $f(x) = 100$, then (4) gives the result stated in (6).

2. (a) Suppose that in place of the homogeneous Dirchlet condition (1c) we have the homogeneous Neumann condition (7). This time use an *even* extension of f and, carrying out steps analogous to those in (3), derive the result

$$u(x,t) = \int_0^\infty f(\xi)\,\frac{e^{-\dfrac{(\xi-x)^2}{4\alpha^2 t}} + e^{-\dfrac{(\xi+x)^2}{4\alpha^2 t}}}{2\alpha\sqrt{\pi t}}\,d\xi. \quad (2.1)$$

(b) Verify that (2.1) does satisfy (7).
(c) Use (2.1) to determine $u(x,t)$ for the case where $f(x) = 100$.

3. Prove that (11) does imply (12).

4. Prove item 4, that if $F(x)$ is even then $F'(0) = 0$.

5. Prove item 6, that if $F(x)$ is even, then $F(x), F''(x), F''''(x), \dots$ are even and $F'(x), F'''(x), \dots$ are odd.

6. Prove item 7, that if $F(x)$ is odd, then $F(x), F''(x), F''''(x), \dots$ are odd and $F'(x), F'''(x), \dots$ are even.

7. (a) Prove that if $F(x,t)$ is an even function of x, then so is $F_t(x,t)$.
(b) Prove that if $F(x,t)$ is an odd function of x, then so is $F_t(x,t)$.

8. State whether or not L is of the linear and even form (19), and briefly state your reasoning.

(a) $L[u] = u_{xx} - u_t - 3u$
(b) $L[u] = u_{xx} + e^{-x}u_t$
(c) $L[u] = u_{xx} + u_{tt} + u_{xt} + u_x + u_t + u$
(d) $L[u] = u_{xx} - u_t + u^2$
(e) $L[u] = u_{xx} + (\sin x)u_x - u_{tt}$
(f) $L[u] = u_{xx} + (\cos x)u_x - 6u_t$
(g) $L[u] = u_{xx} - u_{tt} + 2u$
(h) $L[u] = u_{xx} + u_{tt} - u_t + 5u$
(i) $L[u] = u_{xx} - u_{tt} - (3\cos 5x)u_t$
(j) $L[u] = u_{xx} - (\sin t)u_t$
(k) $L[u] = u_{xx} + xtu_x - u_t$
(l) $L[u] = (x^2u_x)_x - u_t$
(m) $L[u] = (x^3u_x)_x - u_{tt} + u$

18.6 Numerical Solution

18.6.1. The finite-difference method. We have seen that a wide variety of diffusion problems can be solved analytically using separation of variables or an integral transform. In more complicated cases one often gives up the hope of obtaining analytical solutions and turns, instead, to a numerical solution technique. For instance, the problem

$$\alpha^2 u_{xx} = u_t, \qquad (0 < x < L,\ 0 < t < \infty) \quad (1a)$$
$$u(0,t) = p(t), \qquad (0 < t < \infty) \quad (1b)$$

$$u(L,t) = q(t), \qquad (0 < t < \infty) \tag{1c}$$
$$u(x,0) = f(x) \qquad (0 < x < L) \tag{1d}$$

is readily solved by separation of variables if $p(t)$ and $q(t)$ are constants, but is so much more difficult if they are not constants (see Exercise 20, Section 18.3) that we might very well turn to numerical solution instead.

In this section we introduce one of the most important techniques for the numerical solution of PDE's, the **finite-difference method**. To explain that method, let us use the representative problem (1).

Our first step is to discretize the problem so that we seek $u(x,t)$ not over the entire x,t domain but only at discrete *grid points* or *nodal points*, with coordinates x_j, t_k in the x,t plane. That is, we divide L into N equal parts, of length $\Delta x = L/N$, and define $x_j = j\Delta x$, for $j = 0, 1, \ldots, N$. Further, we choose a time increment Δt and define $t_k = k\Delta t$, for $k = 0, 1, 2, \ldots$. The resulting set of grid points, known as the *computational grid*, is shown in Fig. 1. At the open circle points u is known, from the initial or boundary conditions, and we wish to solve for u at the solid circle points.

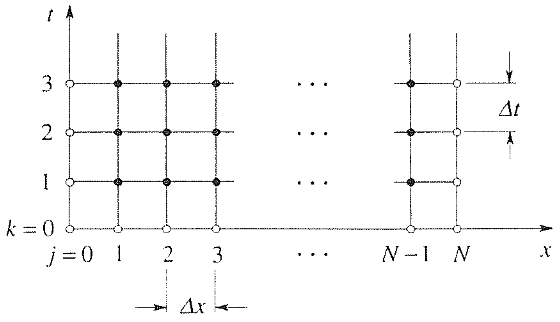


Figure 1. The computational grid.

Next, we seek a finite-difference approximation of the PDE (1a) that will relate u at the various grid points. For the u_t term in (1a) we use the difference quotient approximation

$$u_t(x,t) = \lim_{\Delta t \to 0} \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} \approx \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t}. \tag{2}$$

That is, we do not take the limit as $\Delta t \to 0$; we choose a small Δt and accept the error that results. We can treat the u_{xx} term in the same manner if we deal with one derivative at a time. Accordingly, write

$$u_{xx} = (u_x)_x \approx \frac{u_x(x+\Delta x,t) - u_x(x,t)}{\Delta x} \tag{3}$$

and then approximate each of the derivatives in the numerator in the same way. Thus,

$$u_{xx}(x,t) \approx \frac{\dfrac{u(x+\Delta x,t) - u(x,t)}{\Delta x} - \dfrac{u(x,t) - u(x-\Delta x,t)}{\Delta x}}{\Delta x} \tag{4}$$

or[*]

$$u_{xx}(x,t) \approx \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{(\Delta x)^2}. \tag{5}$$

Putting (2) and (5) into (1a), with $x = x_j$, $x + \Delta x = x_{j+1}$, $x - \Delta x = x_{j-1}$, $t = t_k$, and $t + \Delta t = t_{k+1}$, gives

$$\alpha^2 \frac{u(x_{j+1},t_k) - 2u(x_j,t_k) + u(x_{j-1},t_k)}{(\Delta x)^2} \approx \frac{u(x_j,t_{k+1}) - u(x_j,t_k)}{\Delta t} \tag{6}$$

or

$$\alpha^2 \frac{U_{j-1,k} - 2U_{j,k} + U_{j+1,k}}{(\Delta x)^2} = \frac{U_{j,k+1} - U_{j,k}}{\Delta t} \tag{7}$$

as our **finite-difference approximation** of the PDE (1a). Although not essential, we distinguish $u(x_j, t_k)$ and U_{jk} as follows: $u(x_j, t_k)$ is the exact solution of the PDE (1a) at x_j, t_k, whereas U_{jk} is the exact solution of the difference equation (7) at x_j, t_k. Since the two are not (in general) identical, because of the difference quotient approximations (2) and (5), it is useful to denote them by different letters, as we have.

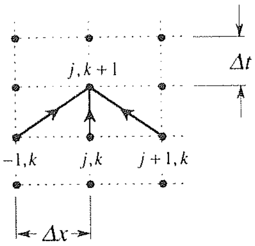


Figure 2. Schematic of (8).

If we solve (7) for $U_{j,k+1}$ we obtain[†]

$$\boxed{U_{j,k+1} = rU_{j-1,k} + (1-2r)U_{j,k} + rU_{j+1,k},} \tag{8}$$

where

$$r \equiv \alpha^2 \frac{\Delta t}{(\Delta x)^2}. \tag{9}$$

Equation (8) enables us to compute U at a given grid point as a linear combination of U's at the preceding time as indicated by the arrows in Fig. 2. Thus, it provides us with a "marching scheme" with which we can march out a solution, one line at a time, subject to the initial and boundary conditions

$$\begin{aligned} U_{j,0} &= f(j\Delta x) \equiv f_j, & (j = 1, 2, \dots, N-1) \qquad &\text{(10a)} \\ U_{0,k} &= p(k\Delta t) \equiv p_k, & (k = 1, 2, \dots) \qquad &\text{(10b)} \\ U_{N,k} &= q(k\Delta t) \equiv q_k. & (k = 1, 2, \dots) \qquad &\text{(10c)} \end{aligned}$$

That is, beginning at the initial time $k = 0$ we can use (8), together with (10), to compute the U's all along the line $k = 1$, then along the line $k = 2$, and so on.

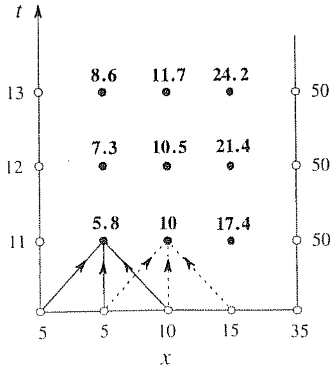


Figure 3. Example 1.

EXAMPLE 1. In (1), let $\alpha^2 = 1$, $L = 1$, $p(t) = 10 + 100t$, $q(t) = 50$, and $f(x) = 20x$. For purposes of illustration, let it suffice to take $N = 4$ (so $\Delta x = 0.25$) and $\Delta t = 0.01$. Then the grid and the initial and boundary values are as shown in Fig. 3. Where did

[*] We use a *forward* difference quotient (forward and backward difference quotients are defined in Section 16.4) in (3) and then *backward* difference quotients in the numerator of the right-hand side of (4) so that the result (5) is a *centered* formula; i.e., centered about x. These choices are discussed further in Exercise 1.

[†] The finite-difference scheme (8) is generally attributed to *E. Schmidt* (1924) and *L. Binder* (1911).

we get the value $u(0,0) = 5$ in the figure? The data is discontinuous at that point because $u(x,0) = f(x) = 20x = 0$ there, whereas $u(0,t) = 10+100t = 10$ there. Consistent with the fact that diffusion is a smoothing process, it seems reasonable to use the average value $(0+10)/2 = 5$ there. Similarly, we use the average value $u(1,0) = (20+50)/2 = 35$ at the other corner. From (9), $r = (1)(0.01)/(0.25)^2 = 0.16$, so (8) becomes

$$U_{j,k+1} = 0.16U_{j-1,k} + 0.68U_{j,k} + 0.16U_{j+1,k}. \tag{11}$$

Sweeping across the first time line, (11) gives

$$\begin{aligned} U_{1,1} &= 0.16U_{0,0} + 0.68U_{1,0} + 0.16U_{2,0} \\ &= 0.16(5) + 0.68(5) + 0.16(10) = 5.8, \\ U_{2,1} &= 0.16U_{1,0} + 0.68U_{2,0} + 0.16U_{3,0} \\ &= 0.16(5) + 0.68(10) + 0.16(15) = 10, \\ U_{3,1} &= 0.16U_{2,0} + 0.68U_{3,0} + 0.16U_{4,0} \\ &= 0.16(10) + 0.68(15) + 0.16(35) = 17.4. \end{aligned} \tag{12}$$

Moving up to the second time line, (11) gives

$$\begin{aligned} U_{1,2} &= 0.16(11) + 0.68(5.8) + 0.16(10) = 7.3, \\ U_{2,2} &= 0.16(5.8) + 0.68(10) + 0.16(17.4) = 10.5, \\ U_{3,2} &= 0.16(10) + 0.68(17.4) + 0.16(50) = 21.4, \end{aligned} \tag{13}$$

and so on.

COMMENT. Since the difference approximations (2) and (5) become exact only as $\Delta t \to 0$ and $\Delta x \to 0$, respectively, it is clear that we need Δt and Δx to be sufficiently small if we are to expect accurate results. Are $\Delta t = 0.01$ and $\Delta x = 0.25$ sufficiently small? Remember that smallness (or largeness) is always *relative* to some reference. Although there is no ready-made reference time with which to compare Δt, observe from the PDE (1a) that α^2/L^2 has the dimensions of 1/time. Thus, we can use $T = L^2/\alpha^2$ as a reference time. And as a reference length we can simply use L. Consequently, for our results to be accurate we need both

$$\frac{\Delta t}{T} = \frac{\Delta t}{L^2/\alpha^2} \ll 1 \qquad \text{and} \qquad \frac{\Delta x}{L} \ll 1 \tag{14}$$

or, since $\alpha = L = 1$ in this example,

$$\Delta t \ll 1 \qquad \text{and} \qquad \Delta x \ll 1. \tag{15}$$

Although $\Delta t = 0.01$ is much smaller than unity, $\Delta x = 0.25$ is not, so we can expect our results to provide only a rough approximation of the exact solution. However, realize that (14) is only a rule of thumb. Typically, one runs the calculation, then reduces Δt and Δx and runs it again, until a sufficient degree of convergence is achieved. ▮

Example 1 illustrates the numerical implementation of (8). In Example 2 let us test the method by choosing Δx and Δt that easily satisfy (15) and comparing the numerical results with a known exact solution.

EXAMPLE 2. *Test Case.* Let us apply (8) to the case where $\alpha^2 = 1$, $L = 1$, $p(t) = q(t) = 0$, and $f(x) = 100$ because this case can be solved exactly by separation of variables [since $p(t)$ and $q(t)$ are constants]. Specifically, we have the exact solution

$$u(x,t) = \frac{400}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin n\pi x \; e^{-(n\pi)^2 t} \tag{16}$$

with which to compare our numerical solution. With (15) in mind, let us choose $\Delta x = 0.02$ (i.e., $N = 50$) and $\Delta t = 0.00018$, in which case $r = 0.00018/(0.02)^2 = 0.45$.

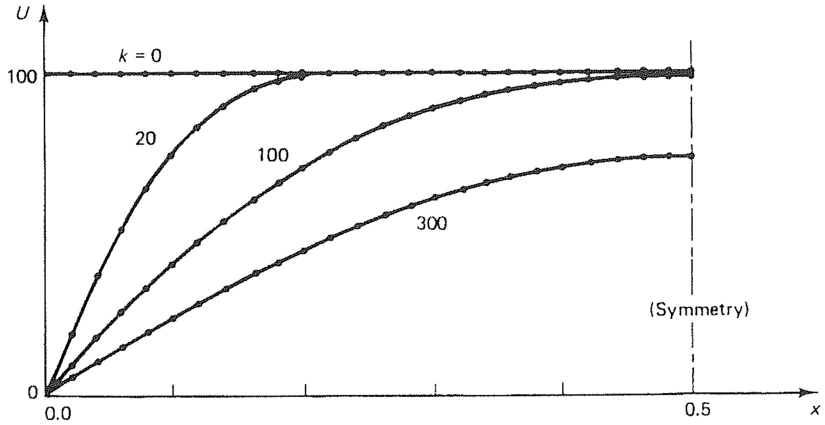


Figure 4. Application of (8), for the case $p(t) = q(t) = 0$, $f(x) = 100$, $L = 1$, $\alpha^2 = 1$, $\Delta x = 0.02$, $\Delta t = 0.00018$ ($r = 0.45$).

In this case our computational grid is quite fine so there are too many calculations to do by hand. However, it is easy to program the calculation (11) with a "do loop" on k and with the initial and boundary conditions

$$U_{j,0} = 100, \qquad (j = 1, 2, \ldots, 49) \tag{17a}$$
$$U_{0,k} = 0, \qquad (k = 1, 2, \ldots) \tag{17b}$$
$$U_{50,k} = 0. \qquad (k = 1, 2, \ldots) \tag{17c}$$

For the corner points (where the data are discontinuous) use average values as in Example 1: $U_{0,0} = 50$, $U_{50,0} = 50$. The numerical results are shown as dots in Fig. 4, and the solid curves correspond to the exact solution given by (16). We have plotted the results only for $k = 20, 100$, and 300 and only over $0 < x < 0.5$ because the solution is symmetric about the midpoint $x = 0.5$. Although the exact and numerical results are not identical, they can hardly be distinguished in Fig. 4. ▮

The results of the test case in Example 2 are encouraging. In fact, if we change the boundary temperatures $p(t)$ and $q(t)$ from constants to time-varying functions, then the analytical separation of variables solution is made much more difficult, yet the numerical finite-difference solution based upon equation (8) is unchanged, except for the data on the right-hand sides of (17b) and (17c).

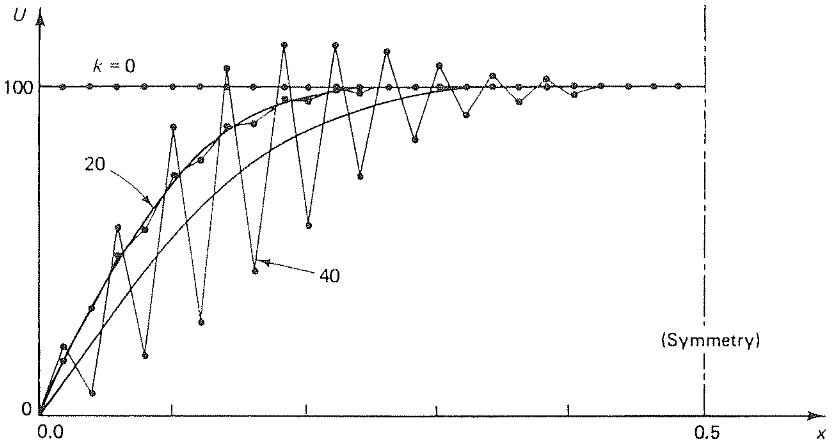


Figure 5. Example 2, with Δt increased to 0.00022.

However, we must qualify our endorsement of the finite difference scheme given by equation (8). For suppose we change Δt even slightly, from 0.00018 to 0.00022. We see from Fig. 5 that the results quickly degenerate: after only 20 time steps a deviation from the exact solution is apparent, and by the time k = 40 the results are worthless. The error is oscillatory, both spatially (with period $2\Delta x$) and temporally (with period $2\Delta t$, although that fact is not observable in the figure because the plots are not for consecutive k values). Even if we do not rely on the exact solution for comparison (indeed, in real applications we do not know the exact solution) it seems clear that the oscillations are some sort of numerical instability rather than a faithful representation of a physical reality because it would surely be unlikely that the spatial and temporal periodicity of such a physical event would exactly equal $2\Delta x$ and $2\Delta t$, respectively. To test this assertion we can halve Δx, say, and rerun the calculation. Sure enough, an oscillation will result once again, this time with spatial period $2\Delta x'$, where $\Delta x' = \Delta x/2$ is the new spacing. Furthermore, we know that diffusion is a smoothing process, whereas the numerical results in Fig. 5 reveal quite the opposite tendency. Thus, even if we do not have the exact solution for comparison, it is clear that the oscillations imply some sort of numerical instability.

Short of a detailed analysis, let us briefly explain the breakdown observed in Fig. 5. Recall that we have used different letters to distinguish the exact solution $u(x,t)$ from the approximate solution $U_{j,k}$ generated by the finite-difference equation (8). We call the difference $u(x_j, t_k) - U_{j,k}$ the **accumulated truncation error** at the j, k grid point, namely, the error incurred by replacing u_t and u_{xx} in (1a) by the finite-difference approximations (2) and (5), respectively. Besides the accumulated truncation error there is an additional error called the **accumulated roundoff error**, incurred because the computer rounds off numbers after a finite number of significant figures. Thus, if we further distinguish $U_{j,k}$ as the values computed by a "perfect computer" (one that keeps an infinite number of significant figures), and

$U^*_{j,k}$ as the actual printout of the real computer, then we can express the total error as

$$\begin{aligned}\text{total error} &= u(x_j, t_k) - U^*_{j,k} \\ &= [u(x_j, t_k) - U_{j,k}] + [U_{j,k} - U^*_{j,k}] \\ &= [\text{accumulated truncation error}] \\ &\quad +[\text{accumulated roundoff error}]. \end{aligned} \tag{18}$$

Regarding the accumulated truncation error, two closely related questions come to mind.

1. With Δx and Δt fixed, what is the behavior of the accumulated truncation error as $k \to \infty$?

2. At the fixed points in the x, t domain, does the accumulated truncation error tend to zero as the mesh is continually refined?

Let us rephrase the second question, which is crucial: At any chosen fixed point in the x, t domain, is it possible to reduce the accumulated truncation error to be smaller in magnitude than any prescribed number by sufficiently refining the grid, that is, by sufficiently reducing Δx and Δt? If so, we say that the finite-difference scheme is **convergent**.

Since the roundoff error enters randomly, we simply ask that the accumulated roundoff error remain small – for instance, that it remain bounded as $k \to \infty$. If so, we say that the scheme is **stable**. (Be aware that these definitions are not entirely standard from one text to another. At least, our terminology here is consistent with the terminology used in our analogous discussion for ODE's in Section 6.5.2.)

Analysis reveals that the finite-difference method (8) is both convergent and stable if Δx and Δt satisfy the criterion

$$\boxed{r = \alpha^2 \frac{\Delta t}{(\Delta x)^2} \le \frac{1}{2},} \tag{19}$$

and is both divergent and unstable if $r > \frac{1}{2}$. [Sure enough, the excellent results displayed in Fig. 4 correspond to $r = (1)(0.00018)/(0.02)^2 = 0.45$, which satisfies (19), whereas those displayed in Fig. 5 correspond to $r = (1)(0.00022)/(0.02)^2 = 0.55$, which does not.] The analysis behind (19) is beyond our present scope,[†] but we do outline the stability part of the analysis in the exercises and urge you to study the latter because it is a typical and powerful application of the matrix eigenvalue problem to the analysis of finite-difference methods.

The restriction (19) may be quite severe, for if Δx is chosen small for the sake of accuracy, then the maximum Δt allowed by (19) may be so small that it necessitates a great many time steps and, consequently, considerable computer time. In

[†] See, for instance, G. D. Smith, *Numerical Solution of Partial Differential Equations* (New York: Oxford University Press, 1965) or R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed. (New York: Interscience, 1967).

the optional remainder of this section we show how to modify the method so as to relieve us of the restriction (19).

18.6.2. Implicit methods: Crank–Nicolson, with iterative solution. (Optional) Equation (8) is by no means the *only* possible finite-difference method for the diffusion equation (1a). For example, the x-wise differencing in (4) is at the initial time t, whereas some weighted average over the time interval (t to $t + \Delta t$) should be more accurate, namely,

$$\begin{aligned} u_{xx}(x,t) \approx (1-\theta)\, &\frac{u(x+\Delta x, t) - 2u(x,t) + u(x-\Delta x, t)}{(\Delta x)^2} \\ &+\theta\, \frac{u(x+\Delta x, t+\Delta t) - 2u(x,t+\Delta t) + u(x-\Delta x, t+\Delta t)}{(\Delta x)^2}, \end{aligned} \tag{20}$$

where the number θ is specified so that $0 \le \theta \le 1$. Then, in place of (7) we have

$$\begin{aligned} &\alpha^2 \left[(1-\theta)\, \frac{U_{j-1,k} - 2U_{j,k} + U_{j+1,k}}{(\Delta x)^2} + \theta\, \frac{U_{j-1,k+1} - 2U_{j,k+1} + U_{j+1,k+1}}{(\Delta x)^2} \right] \\ &\quad = \frac{U_{j,k+1} - U_{j,k}}{\Delta t}, \end{aligned} \tag{21}$$

which reduces to (7) if $\theta = 0$. It turns out that the parameter θ gives us the desired control. Specifically, it can be shown that if $\theta \ge \frac{1}{2}$, then (21) is convergent and stable for *all* $r > 0$; that is, if we use any θ greater than or equal to 1/2, then the condition (19) can be discarded. The borderline case $\theta = \frac{1}{2}$ gives the well-known **Crank–Nicolson** scheme

$$\boxed{\begin{aligned} -rU_{j-1,k+1} + 2(1+r)U_{j,k+1} - rU_{j+1,k+1}& \\ = rU_{j-1,k} + 2(1-r)U_{j,k} + rU_{j+1,k}& \end{aligned}} \tag{22}$$

for $j = 1, 2, \dots, N-1$ and $k = 0, 1, 2, \dots$, where $r = \alpha^2 \Delta t/(\Delta x)^2$, as before.

Given the initial and boundary values, we can use (22) to compute the first line of unknowns $U_{1,1}, U_{2,1}, \dots, U_{N-1,1}$, then the second line of unknowns $U_{1,2}, U_{2,2}, \dots, U_{N-1,2}$, and so on, just as we did using (8). However, (22) presents a difficulty which we illustrate by writing it out for the first line of unknowns, with $r = 1$ and $N = 5$, say, for definiteness.

$$\begin{aligned} j = 1: &\quad -\underline{U_{0,1}} + 4U_{1,1} - U_{2,1} = U_{0,0} + U_{2,0}, \\ j = 2: &\quad -U_{1,1} + 4U_{2,1} - U_{3,1} = U_{1,0} + U_{3,0}, \\ j = 3: &\quad -U_{2,1} + 4U_{3,1} - U_{4,1} = U_{2,0} + U_{4,0}, \\ j = 4: &\quad -U_{3,1} + 4U_{4,1} - \underline{U_{5,1}} = U_{3,0} + U_{5,0}. \end{aligned} \tag{23}$$

The two underlined terms are known boundary values, so let us move them to the right with the other known terms. Then (23) becomes

$$\begin{aligned} 4U_{1,1} - U_{2,1} \qquad\qquad\qquad &= U_{0,1} + U_{0,0} + U_{2,0}, \\ -U_{1,1} + 4U_{2,1} - U_{3,1} \qquad\quad &= U_{1,0} + U_{3,0}, \\ - U_{2,1} + 4U_{3,1} - U_{4,1} &= U_{2,0} + U_{4,0}, \\ - U_{3,1} + 4U_{4,1} &= U_{3,0} + U_{5,0} + U_{3,1}. \end{aligned} \tag{24}$$

The difficulty, which we can see clearly in (24), is that the equations (24) are *coupled*. Thus, we need to solve the matrix equation (24) for the first line of unknowns $U_{1,1}, U_{2,1}, U_{3,1}, U_{4,1}$, then increment k by 1 in (22) and solve for a similar matrix equation for the second line of unknowns $U_{1,2}, U_{2,2}, U_{3,2}, U_{4,2}$, and so on. In contrast, the scheme (8) gives *uncoupled* equations for $U_{1,k+1}, U_{2,k+1}, \ldots, U_{N-1,k+1}$. Thus, we call (22) an **implicit** scheme, whereas (8) is an **explicit** scheme. Graphically, the Crank–Nicolson scheme (22) corresponds to the computational pattern shown in Fig. 6, which reveals the *nearest-neighbor coupling* and is in contrast with the pattern shown in Fig. 2 for the scheme (8).

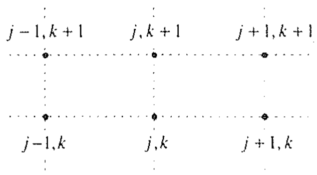


Figure 6. Implicit scheme.

Expressing (22) in matrix form and recalling the initial and boundary conditions given by (10) gives

$$\begin{bmatrix} 2(1+r) & -r & & \cdots & 0 \\ -r & 2(1+r) & -r & & \vdots \\ & & \ddots & & \\ \vdots & & -r & 2(1+r) & -r \\ 0 & \cdots & & -r & 2(1+r) \end{bmatrix} \begin{bmatrix} U_{1,k+1} \\ U_{2,k+1} \\ \vdots \\ U_{N-2,k+1} \\ U_{N-1,k+1} \end{bmatrix}$$
$$= \begin{bmatrix} rp_{k+1} + rp_k + 2(1-r)U_{1,k} + rU_{2,k} \\ rU_{1,k} + 2(1-r)U_{2,k} + rU_{3,k} \\ \vdots \\ rU_{N-3,k} + 2(1-r)U_{N-2,k} + rU_{N-1,k} \\ rU_{N-2,k} + 2(1-r)U_{N-1,k} + rq_{k+1} + rq_k \end{bmatrix}. \tag{25}$$

The system (25) is of the matrix form

$$\mathbf{A}\mathbf{U}_{k+1} = \mathbf{c}, \tag{26}$$

where $\mathbf{A}$ is *tridiagonal* (due to the nearest-neighbor coupling) and *symmetric*. $\mathbf{A}$ and $\mathbf{c}$ are known and $\mathbf{U}_{k+1}$ is the unknown. The idea is to set $k = 0$ and solve (25) for the entire line of unknowns $U_{1,1}, U_{2,1}, \ldots, U_{N-1,1}$, then set $k = 1$ and solve for the line $U_{1,2}, U_{2,2}, \ldots, U_{N-1,2}$, and so on.

Notice carefully that the situation is not as bad as it may appear to be because $\mathbf{A}$ is strongly diagonal. That is, the off-diagonal terms ($-r$'s and 0's) are small compared to the diagonal terms [$2(1+r)$'s], so (25) is *almost* uncoupled. Thus, we say that (25) is only *weakly coupled*. To take advantage of this circumstance, let us split $\mathbf{A}$ into its diagonal part plus the deviation from that,

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} 2(1+r) & 0 & & \cdots & 0 \\ 0 & 2(1+r) & & & \vdots \\ & & \ddots & & \\ \vdots & & & 2(1+r) & \\ 0 & \cdots & & 0 & 2(1+r) \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & -r & 0 & & \cdots & 0 \\ -r & 0 & -r & & & \vdots \\ & & & \ddots & & \\ \vdots & & & -r & 0 & -r \\ 0 & \cdots & & 0 & -r & 0 \end{bmatrix} \\ &\equiv 2(1+r)\mathbf{I} + \mathbf{A}', \end{aligned} \tag{27}$$

where $\mathbf{I}$ is an $(N-1) \times (N-1)$ identity matrix and $\mathbf{A}'$ is the "deviation" matrix. Then (26) becomes

$$[2(1+r)\mathbf{I} + \mathbf{A}']\mathbf{U}_{k+1} = \mathbf{c} \tag{28}$$

or

$$\mathbf{U}_{k+1} = \frac{1}{2(1+r)}\,\mathbf{c} - \frac{1}{2(1+r)}\,\mathbf{A}'\mathbf{U}_{k+1}. \tag{29}$$

Since $\mathbf{A}'$ is small compared to $2(1+r)\mathbf{I}$, in (28), we could neglect it altogether and obtain, from (29),

$$\mathbf{U}_{k+1} \approx \frac{1}{2(1+r)}\,\mathbf{c}. \tag{30}$$

Better yet, we could accept (30) as an initial approximation

$$\mathbf{U}_{k+1}^{(0)} = \frac{1}{2(1+r)}\,\mathbf{c} \tag{31}$$

and obtain an improved result as

$$\mathbf{U}_{k+1}^{(1)} = \frac{1}{2(1+r)}\,\mathbf{c} - \frac{1}{2(1+r)}\,\mathbf{A}'\mathbf{U}_{k+1}^{(0)}. \tag{32}$$

In fact, repeating this procedure, we have the **iterative** algorithm

$$\boxed{\mathbf{U}_{k+1}^{(n+1)} = \frac{1}{2(1+r)}\left[\mathbf{c} - \mathbf{A}'\mathbf{U}_{k+1}^{(n)}\right],} \tag{33}$$

where $\mathbf{U}_{k+1}^{(n+1)}$ is the $(n+1)$st iterate of $\mathbf{U}_{k+1}$. With k fixed, we carry out (33) for $n = 0, 1, 2, \ldots$ [where $\mathbf{U}_{k+1}^{(0)}$ is given by (31)], until suitable convergence is attained. For example, we might continue the iteration until each element of $\mathbf{U}_{k+1}^{(n+1)}$ and the corresponding element of $\mathbf{U}_{k+1}^{(n)}$ differ in magnitude by less than 10^{-5}.

The scheme given by (33) and (31) is called **Jacobi iteration**, and it is shown in the exercises that Jacobi iteration converges to the exact solution of $\mathbf{A}\mathbf{U}_{k+1} = \mathbf{c}$ for all finite values of r. In fact, we can improve upon (33) slightly, by using the $(n+1)$st components as soon as they become available. To elaborate, let us write out the Jacobi scheme (33):

$$\begin{aligned}
U_{1,k+1}^{(n+1)} &= \frac{1}{2(1+r)}\left[c_1 - a'_{1,1}U_{1,k+1}^{(n)} - \cdots - a'_{1,N-1}U_{N-1,k+1}^{(n)}\right] \\
U_{2,k+1}^{(n+1)} &= \frac{1}{2(1+r)}\left[c_2 - a'_{2,1}U_{1,k+1}^{(n)} - a'_{2,2}U_{2,k+1}^{(n)}\right. \\
&\qquad\qquad \left. - \cdots - a'_{2,N-1}U_{N-1,k+1}^{(n)}\right] \\
&\ \ \vdots \\
U_{N-1,k+1}^{(n+1)} &= \frac{1}{2(1+r)}\left[c_{N-1} - a'_{N-1,1}U_{1,k+1}^{(n)}\right. \\
&\qquad \left. - \cdots - a'_{N-1,N-2}U_{N-2,k+1}^{(n)} - a'_{N-1,N-1}U_{N-1,k+1}^{(n)}\right].
\end{aligned} \tag{34}$$

Having computed $U_{1,k+1}^{(n+1)}$ from the first equation in (34), let us use that value in place of the less-up-to-date value $U_{1,k+1}^{(n)}$ that appears in the right-hand side of the second equation. Similarly, let us use the already computed values $U_{1,k+1}^{(n+1)}, U_{2,k+1}^{(n+1)}$ in place of $U_{1,k+1}^{(n)}, U_{2,k+1}^{(n)}$ in the right-hand side of the third equation, and so on. This idea produces the **Gauss–Seidel** scheme

$$\boxed{\mathbf{U}_{k+1}^{(n+1)} = \frac{1}{2(1+r)}\left[\mathbf{c} - \mathbf{L}\mathbf{U}_{k+1}^{(n+1)} - \mathbf{M}\mathbf{U}_{k+1}^{(n)}\right],} \tag{35}$$

where $\mathbf{L}$ and $\mathbf{M}$ are the lower and upper parts, respectively, of $\mathbf{A}'$. That is, $\mathbf{A}' = \mathbf{L} + \mathbf{M}$, where

$$\mathbf{L} = \begin{bmatrix} 0 & & & \cdots & 0 \\ -r & 0 & & & \vdots \\ 0 & -r & 0 & & \\ \vdots & & & \ddots & \\ 0 & \cdots & 0 & -r & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 & -r & 0 & \cdots & 0 \\ & 0 & -r & & \vdots \\ & & 0 & & \\ \vdots & & & \ddots & -r \\ 0 & \cdots & & & 0 \end{bmatrix}. \tag{36}$$

Like the Jacobi method, the Gauss–Seidel method converges for all finite r, and approximately twice as fast.

Finally, there is another simple improvement that will further increase the speed of convergence. Re-expressing (35) as

$$\begin{aligned}\mathbf{U}_{k+1}^{(n+1)} &= \mathbf{U}_{k+1}^{(n)} + \frac{1}{2(1+r)}\left\{\mathbf{c} - \mathbf{L}\mathbf{U}_{k+1}^{(n+1)} - [\mathbf{M} + 2(1+r)\mathbf{I}]\mathbf{U}_{k+1}^{(n)}\right\} \\ &\equiv \mathbf{U}_{k+1}^{(n)} + \mathbf{\Delta U}_{k+1}^{(n)}, \end{aligned} \tag{37}$$

we insert a numerical "control parameter" ω:

$$\mathbf{U}_{k+1}^{(n+1)} = \mathbf{U}_{k+1}^{(n)} + \omega\mathbf{\Delta U}_{k+1}^{(n)}. \tag{38}$$

The idea is general and can be applied to any iterative scheme: $\mathbf{\Delta U}_{k+1}^{(n)}$ is a "correction term," and adjusting the size of the correction, by means of ω, might speed the convergence. Often, one chooses ω based on numerical experimentation, but in the present case it can be shown analytically that the optimum ω is given by

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\mu^2}}, \qquad \text{where} \qquad \mu = \frac{r}{1+r}\cos\frac{\pi}{N}. \tag{39}$$

We see from (39) that ω_{opt} lies somewhere between 1 and 2. Since $\omega_{\text{opt}} > 1$, the modified Gauss–Seidel scheme (38) is known as **successive overrelaxation**, or **SOR** for brevity.

EXAMPLE 3. To illustrate the Jacobi, Gauss–Seidel, and SOR methods, consider the problem where $u(0,t) = u(L,t) = 0$ and $u(x,0) = 100$, and choose $N = 4$ and $r = 1$. Then the Crank–Nicolson scheme (22) gives these simultaneous equations for the unknown values $U_{1,1}$, $U_{2,1}$, $U_{3,1}$ in the first "time-line":

$$\begin{aligned} 4U_{1,1} - U_{2,1} &= 150, \\ -U_{1,1} + 4U_{2,1} - U_{3,1} &= 200, \\ -U_{2,1} + 4U_{3,1} &= 60. \end{aligned} \tag{40}$$

The latter is readily solved, and we obtain

$$U_{1,1} = 55.54, \quad U_{2,1} = 72.14, \quad U_{3,1} = 33.04, \tag{41}$$

but let us use (40) to illustrate the three methods of iterative solution.

Jacobi:

$$\begin{aligned} U_{1,1}^{(n+1)} &= \frac{1}{4}(U_{2,1}^{(n)} + 150), \\ U_{2,1}^{(n+1)} &= \frac{1}{4}(U_{1,1}^{(n)} + U_{3,1}^{(n)} + 200), \\ U_{3,1}^{(n+1)} &= \frac{1}{4}(U_{2,1}^{(n)} + 60), \end{aligned} \tag{42}$$

so

$$\begin{aligned}
U_{1,1}^{(0)} &= \frac{1}{4}(0+150) = \underline{37.5},\\
U_{2,1}^{(0)} &= \frac{1}{4}(0+0+200) = \underline{50},\\
U_{3,1}^{(0)} &= \frac{1}{4}(0+60) = \underline{15},
\end{aligned} \tag{43}$$

$$\begin{aligned}
U_{1,1}^{(1)} &= \frac{1}{4}(50+150) = \underline{50},\\
U_{2,1}^{(1)} &= \frac{1}{4}(37.5+15+200) = \underline{63.13},\\
U_{3,1}^{(1)} &= \frac{1}{4}(50+60) = \underline{27.5},
\end{aligned} \tag{44}$$

and so on.
Gauss–Seidel:

$$\begin{aligned}
U_{1,1}^{(n+1)} &= \frac{1}{4}(U_{2,1}^{(n)}+150),\\
U_{2,1}^{(n+1)} &= \frac{1}{4}(U_{1,1}^{(n+1)}+U_{3,1}^{(n)}+200),\\
U_{3,1}^{(n+1)} &= \frac{1}{4}(U_{2,1}^{(n+1)}+60),
\end{aligned} \tag{45}$$

so

$$\begin{aligned}
U_{1,1}^{(0)} &= \frac{1}{4}(0+150) = \underline{37.5},\\
U_{2,1}^{(0)} &= \frac{1}{4}(0+0+200) = \underline{50},\\
U_{3,1}^{(0)} &= \frac{1}{4}(0+60) = \underline{15},
\end{aligned} \tag{46}$$

$$\begin{aligned}
U_{1,1}^{(1)} &= \frac{1}{4}(50+150) = \underline{50},\\
U_{2,1}^{(1)} &= \frac{1}{4}(50+15+200) = \underline{66.25},\\
U_{3,1}^{(1)} &= \frac{1}{4}(66.25+60) = \underline{31.56},
\end{aligned} \tag{47}$$

and so on.
SOR: First, we need to compute ω_{opt} : $\mu = \frac{1}{2}\cos\frac{\pi}{4} = \sqrt{2}/4$, $\omega_{opt} = 2/(1+\sqrt{1-\mu^2})$ $= 1.03$. As in (43) and (46),

$$U_{1,1}^{(0)} = \underline{37.5}, \quad U_{2,1}^{(0)} = \underline{50}, \quad U_{3,1}^{(0)} = \underline{15}. \tag{48}$$

Next, make a "tentative" Gauss–Seidel step using (45) (with $n = 0$) and the values given by (48):

$$U_{1,1}^{(1)} = 50 \quad \text{so} \quad \Delta U_{1,1}^{(0)} = 50 - 37.5 = 12.5,$$

$$\begin{aligned} U_{2,1}^{(1)} &= 66.25 \quad \text{so} \quad \Delta U_{2,1}^{(0)} = 66.25 - 50 = 16.25, \\ U_{3,1}^{(1)} &= 31.56 \quad \text{so} \quad \Delta U_{3,1}^{(0)} = 31.56 - 15 = 16.56. \end{aligned} \tag{49}$$

Now make the SOR step:

$$\begin{aligned} U_{1,1}^{(1)} &= U_{1,1}^{(0)} + \omega\Delta U_{1,1}^{(0)} = 37.5 + 1.03(12.5) = \underline{50.38}, \\ U_{2,1}^{(1)} &= U_{2,1}^{(0)} + \omega\Delta U_{2,1}^{(0)} = 50 + 1.03(16.25) = \underline{66.74}, \\ U_{3,1}^{(1)} &= U_{3,1}^{(0)} + \omega\Delta U_{3,1}^{(0)} = 15 + 1.03(16.56) = \underline{32.06}. \end{aligned} \tag{50}$$

Carrying out one more iteration (Exercise 16) and comparing the successive $U_{2,1}^{(n)}$ values, which should be representative, gives the results presented in Table 1. ▮

Table 1. Successive $U_{2,1}^{(n)}$ values.

	Jacobi	Gauss–Seidel	SOR
$n = 0$	50	50	50
$n = 1$	63.13	66.25	66.74
$n = 2$	69.38	71.41	71.70
⋮	⋮	⋮	⋮
Exact	72.14	72.14	72.14

Closure. The finite-difference method developed quickly beginning around 1950 when digital computers became widely available, and is one of the most important methods for solving partial differential equations. In this text we discuss it in the present section for the diffusion equation and in Section 20.5 for the Laplace equation. The basic idea is to discretize the problem so that instead of seeking $u(x, t)$ over the given x, t domain we seek u only at a finite set of grid points. The PDE is discretized by replacing the various partial derivations by approximate difference quotients so, in place of the PDE, we end up with linear algebraic equations on the unknown $U_{j,k}$ values at the grid points. There are many possible finite-difference schemes for a given PDE, depending upon the forms chosen for the difference quotients. The simplest one for our diffusion equation is given by (8) and is called an explicit scheme because it gives $U_{j,k+1}$ explicitly in terms of known values. However, this scheme is invalid if $r = \alpha^2\Delta t/(\Delta x)^2$ exceeds 1/2, so that if we choose Δx small, for accuracy, then we may need Δt to be so small that a prohibitively large number of time steps may be required. In the optional second part of this section we turn to implicit schemes to eliminate the $r \leq 1/2$ restriction. The price that we pay for this improvement is that at each time step we need to solve an $(N - 1) \times (N - 1)$ matrix equation for the $N - 1$ values of U along that line. Fortunately, the matrix is strongly diagonal, so that efficient iterative methods can be used for the solution.

EXERCISES 18.6

1. We use a forward difference quotient in (3) and then backward difference quotients in the numerator of the right-hand side of (4).

(a) Show that if we use forward difference quotients in (3) and (4), then we obtain the scheme

$$U_{j,k+1} = (1+r)U_{j,k} - 2rU_{j+1,k} + rU_{j+2,k} \tag{1.1}$$

in place of (8).
(b) Show that if we use backward difference quotients in (3) and (4), then we obtain the scheme

$$U_{j,k+1} = (1+r)U_{j,k} - 2rU_{j-1,k} + rU_{j-2,k}. \tag{1.2}$$

(c) Discuss any advantages or disadvantages that occur to you for the schemes (1.1) and (1.2) in comparison with (8).

2. Continue the hand calculation begun in Example 1. Specifically, determine $U_{1,3}, U_{2,3}, U_{3,3}$, and $U_{1,4}, U_{2,4}, U_{3,4}$.

3. Consider the problem $\alpha^2 u_{xx} = u_t$ $(0 < x < 10,\ 0 < t < \infty)$ with the boundary and initial conditions $u(0,t) = 100$, $u(10,t) = 0$, and $u(x,0) = 0$. With $\Delta x = 2.5$, $\Delta t = 2$, and $\alpha^2 = 1$, use (8) to compute the first three "lines" of $U_{j,k}$'s: the nine values $U_{1,1}$ through $U_{3,3}$.

4. Show that if the PDE (1a) is modified to include a Newton cooling term Hu and a heat source distribution term $F(x,t)$, as

$$\alpha^2 u_{xx} = u_t + Hu - F(x,t), \qquad (0 < x < L,\ 0 < t < \infty) \tag{4.1}$$

then in place of (8) we obtain

$$U_{j,k+1} = rU_{j-1,k} + (1 - 2r - H\Delta t)U_{j,k} + rU_{j+1,k} + F_{j,k}\Delta t, \tag{4.2}$$

where $F_{j,k}$ denotes $F(x_j, t_k)$ and H is a constant.

5. In Exercise 4, let $\alpha^2 = 1$, $L = 1$, $u(0,t) = 0, u(x,0) = 0, u(1,t) = 0$, $H = 0$, and $F(x,t) = 10$. With $\Delta t = 0.02$ and $\Delta x = 0.25$, use (4.2) to compute the first three "lines" of $U_{j,k}$'s, i.e., the nine values $U_{1,1}$ through $U_{3,3}$.

6. Repeat Exercise 5, with these changes: $u(0,t) = 100$, $F(x,t) = 10\sin\pi x$.

7. Repeat Exercise 5, with these changes: $u(x,0) = 100$, $H = 4$, $F(x,t) = 0$.

8. To see the smoothing nature of the diffusion process in a simple numerical example, consider the problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \\ u(x,0) &= 100H(x), \end{aligned}$$

where $H(x)$ is the Heaviside function, and the diffusivity is $\alpha^2 = 0.2$. Use (8), with $\Delta t = 0.5$ and $\Delta x = 1$, to compute the $U_{j,k}$ values for the first three lines; i.e., through $k = 3$. Plot $U_{j,k}$ versus j, for $k = 0, 1, 2, 3$.

9. (a) To see the effect of the condition (19) in a simple calculation, use (8) for the problem

$$\begin{aligned} u_{xx} &= u_t, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \\ u(x,0) &= 100H(x), \end{aligned} \tag{9.1}$$

where $H(x)$ is the Heaviside function. With $\Delta t = 0.2$ and $\Delta x = 1$, compute the $U_{j,k}$ values for the first five lines; i.e., through $k = 5$. Keeping $\Delta x = 1$, repeat the calculation with $\Delta t = 0.4, 0.6$, and 0.8.
(b) Plot your results and discuss them.
(c) For the four different cases ($\Delta t = 0.2, 0.4, 0.6, 0.8$), plot your final temperature distribution (i.e., at $k = 5$) and also the exact solution (which can be found in Section 18.4).

10. In (1) let $L = 12$, $p(t) = 100$, $q(t) = 0$, and $f(x) = 0$, and suppose the rod is comprised of two different materials, with $\alpha^2 = 1.8$ over $0 < x < 6$ and $\alpha^2 = 0.2$ over $6 < x < 12$. With $\Delta t = 0.5$ and $\Delta x = 2$, use (8) to compute the $U_{j,k}$ values for the first four lines; i.e., through $k = 4$. Note that (8) holds over $x < 6$ and over $x > 6$ but not at the junction $x = 6$. There, use the fact that the heat flux crossing $x = 6$ from the left must equal the heat flux crossing $x = 6$ toward the right or, from the Fourier law of heat conduction,

$$K_L \left.\frac{\partial u}{\partial x}\right|_{x=6-} = K_R \left.\frac{\partial u}{\partial x}\right|_{x=6+}, \tag{10.1}$$

where $K_L = 25$ is the thermal conductivity of the material to the left of $x = 6$ and $K_R = 3$ is the thermal conductivity of the material to the right of $x = 6$. You will need to express (10.1) in finite-difference form.

11. The problem

$$\begin{aligned} u_{xx} &= u_t, \qquad (0 < x < 1,\ 0 < t < \infty) \\ u(0,t) &= u(1,t) = 0, \quad u(x,0) = 100\sin\pi x \end{aligned} \tag{11.1}$$

admits the exact one-term solution $u(x,t) = 100(\sin \pi x)\exp(-\pi^2 t)$.

(a) Use (8), with $\Delta t = 0.02$ and $\Delta x = 0.25$ to compute the first three lines of $U_{j,k}$'s (i.e., through $k = 3$) and compare your results with the exact values.
(b) Use (8) to evaluate $u(0.5, 0.06)$, with Δt and Δx sufficiently small so that your value of $u(0.5, 0.06)$ is correct to four significant figures. (Use a computer.)
(c) Use (8) to evaluate (by computer) the $U_{j,k}$'s through $k = 45$, with $\Delta t = 0.0010$ and $\Delta x = 0.05$. Plot both your computed solution and the exact solution at $t = 0.045$ (i.e., at $k = 45$).
(d) Use (8) to evaluate (by computer) the $U_{j,k}$'s through $k = 30$, and $\Delta t = 0.0015$ and $\Delta x = 0.05$. Plot both your computed solution and the exact solution at $t = 0.045$ (i.e., at $k = 30$).
(e) Use (8) to evaluate (by computer) the $U_{j,k}$'s through $k = 15$, with $\Delta t = 0.0030$ and $\Delta x = 0.05$. Plot both your computed solution and the exact solution at $t = 0.045$ (i.e., at $k = 15$). Interpret your results in the light of your results to parts (c) and (d) if, indeed, you worked those parts.

12. (*Use of Taylor series*) To derive the finite-difference approximation (5) we used the classical difference quotient definition of the derivative. Derive (5) using Taylor series instead. HINT: Expand $u(x + \Delta x, t)$ about x:

$$u(x+\Delta x, t) = u(x,t) + u_x(x,t)\Delta x + \frac{1}{2}u_{xx}(x,t)(\Delta x)^2 + \cdots.$$

Similarly, expand $u(x - \Delta x, t)$ about x. Add those two formulas, cut off the series on the right-hand side after the first couple of terms (as an approximation), and solve for $u_{xx}(x,t)$.

13. (*Deriving the stability criterion (19)*) We stated that the finite-difference scheme (8) is both convergent and stable if $r \le \frac{1}{2}$, and is both divergent and unstable if $r > \frac{1}{2}$. Here, we outline a proof of the stability part of that claim and ask you to write out the steps and to supply any missing steps or reasoning. To begin, show that (8) can be expressed in matrix form as

$$[U_{1,k+1}, \dots, U_{N-1,k+1}]^{\mathrm{T}} =$$
$$\begin{bmatrix} 1-2r & r & 0 & \cdots & 0 \\ r & 1-2r & r & & \vdots \\ & & \ddots & & \\ \vdots & & r & 1-2r & r \\ 0 & \cdots & 0 & r & 1-2r \end{bmatrix} \begin{bmatrix} U_{1,k} \\ U_{2,k} \\ \vdots \\ U_{N-2,k} \\ U_{N-1,k} \end{bmatrix}$$

$$+ [rU_{0,k}, 0, \dots, 0, rU_{N,k}]^{\mathrm{T}}, \tag{13.1}$$

where we use the transpose notation to save vertical space. More compactly, express (13.1) as

$$\mathbf{U}_{k+1} = \mathbf{A}\mathbf{U}_k + \mathbf{c}_k. \tag{13.2}$$

In stability analyses it is commonly assumed that roundoff errors occur only along the first line ($k = 0$), say, and then to see whether the errors remain bounded as k increases. Thus, in place of the exact values $\mathbf{U}_0$, $\mathbf{c}_0$, the initial roundoff errors result in the actual values $\mathbf{U}_0^*$, $\mathbf{c}_0^*$. The machine proceeds to compute values $\mathbf{U}_1^*, \mathbf{U}_2^*, \dots$ according to

$$\mathbf{U}_{k+1}^* = \mathbf{A}\mathbf{U}_k^* + \mathbf{c}_k^*, \tag{13.3}$$

rather than exact values $\mathbf{U}_1, \mathbf{U}_2, \dots$ according to (13.2). Show that the roundoff error $\mathbf{e}_k \equiv \mathbf{U}_k - \mathbf{U}_k^*$ propagates according to

$$\mathbf{e}_{k+1} = \mathbf{A}\mathbf{e}_k + \mathbf{b}_k, \tag{13.4}$$

where $\mathbf{b}_k = \mathbf{c}_k - \mathbf{c}_k^* = [re_{0,k}, 0, \dots, 0, re_{N,k}]^{\mathrm{T}}$. From (13.4) show that

$$\mathbf{e}_k = \mathbf{A}^k\mathbf{e}_0 + \mathbf{A}^{k-1}\mathbf{b}_0. \tag{13.5}$$

Show that the $(N-1) \times (N-1)$ matrix $\mathbf{A}$ must have $N-1$ orthogonal eigenvectors, say $\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_{N-1}$, so that $\mathbf{e}_0$ and $\mathbf{b}_0$ can be expressed in the form

$$\begin{aligned} \mathbf{e}_0 &= \alpha_1\mathbf{\Phi}_1 + \cdots + \alpha_{N-1}\mathbf{\Phi}_{N-1}, \\ \mathbf{b}_0 &= \beta_1\mathbf{\Phi}_1 + \cdots + \beta_{N-1}\mathbf{\Phi}_{N-1}. \end{aligned} \tag{13.6,7}$$

Putting these into (13.5), show that

$$\begin{aligned} \mathbf{e}_k = (&\alpha_1\lambda_1^k + \beta_1\lambda_1^{k-1})\mathbf{\Phi}_1 + \cdots \\ &+ (\alpha_{N-1}\lambda_{N-1}^k + \beta_{N-1}\lambda_{N-1}^{k-1})\mathbf{\Phi}_{N-1}, \end{aligned} \tag{13.8}$$

where $\lambda_1, \dots, \lambda_{N-1}$ are the eigenvalues of $\mathbf{A}$. Explain why it follows from (13.8) that for stability it is necessary and sufficient that

$$|\lambda_1| \le 1, \dots, |\lambda_{N-1}| \le 1. \tag{13.9}$$

Since $\mathbf{A}$ is a function of r its λ's are too. Hence, the most restrictive of the conditions (13.9) should give (19). Show that (13.9) does, indeed, give $r \le \frac{1}{2}$ or, more precisely,

$$r \le \frac{1}{1-\cos\left(\dfrac{N-1}{N}\pi\right)}. \tag{13.10}$$

For large N the right-hand side is approximately $\frac{1}{2}$; e.g., for $N = 50$ (13.10) gives $r \le 0.50049$. HINT: The $\mathbf{A}$ matrix is tridiagonal, of the type shown in Exercise 7 of Section 11.2 with "a" $= r$, "b" $= 1-2r$, and "c" $= r$. Thus, its eigenvalues are, according to equation (7.1) of that exercise,

$$\lambda_n = 1 - 2r + 2r\cos\frac{n\pi}{N} \tag{13.11}$$

for $n = 1, 2, \dots, N-1$. Further, note that each inequality $|\lambda_j| \le 1$ in (13.9) amounts to the statement $-1 \le \lambda_j \le 1$, hence the *two* inequalities $-1 \le \lambda_j$ and $\lambda_j \le 1$. Of the $2N-2$ inequalities in (13.9), you should find that the most restrictive condition on r is given by (13.10).

14. (*Stability of implicit scheme*) The stability of the implicit scheme (21) for all r (if $\theta \ge \frac{1}{2}$) is a striking result. Proof of that result would follow the same lines as the proof that is outlined in Exercise 13 for the explicit scheme. However, for pedagogical purposes it might be better to consider the simpler case of an *ordinary* differential equation. Specifically, consider the simple *test equation*

$$u_t = -Au, \tag{14.1}$$

where A is a prescribed positive constant, with the implicit finite-difference approximation

$$\frac{U_{k+1} - U_k}{\Delta t} = -(1-\theta)AU_k - \theta AU_{k+1}. \tag{14.2}$$

Following the same lines as in Exercise 13, show that the roundoff error $e_k \equiv U_k - U_k^*$ propagates according to

$$e_{k+1} = Ke_k, \tag{14.3}$$

where

$$K = \frac{1 - A(1-\theta)\Delta t}{1 + A\theta\Delta t}. \tag{14.4}$$

Thus, show that the scheme is stable if and only if

$$A(1-2\theta)\Delta t \le 2. \tag{14.5}$$

NOTE: Suppose that $\theta = 0$ so the scheme (14.2) is explicit. Then (14.5) tells us that for stability we must choose $\Delta t \le 2/A$. In some problems, such as occur in the study of chemical kinetics, A can be extremely large, so that $\Delta t \le 2/A$ forces us to use extremely small time steps. However, observe that if we use the implicit scheme (14.2) with $\theta \ge \frac{1}{2}$, then (14.5) is satisfied with no restriction on Δt.

15. Use the Crank–Nicolson scheme (22) to solve the problem (11.1) in Exercise 11 through the first three lines of $U_{j,k}$'s (i.e., through $k = 3$), with $\Delta t = 0.1$ and $\Delta x = 0.25$, using computer software (such as the *Maple* linsolve command) to solve the matrix equation obtained at each time step. Compare your results at $t = 0.3$ with the exact solution.

16. In Example 3 we used the Jacobi, Gauss–Seidel, and SOR methods to work out the iterates $U_{1,1}^{(0)}$, $U_{2,1}^{(0)}$, $U_{3,1}^{(0)}$ and $U_{1,1}^{(1)}$, $U_{2,1}^{(1)}$, $U_{3,1}^{(1)}$. Continuing the calculation, work out $U_{1,1}^{(2)}$, $U_{2,1}^{(2)}$, $U_{3,1}^{(2)}$. NOTE: For $U_{2,1}^{(2)}$ your three values should agree with the values 69.38, 71.41, 71.70 given in Table 1.

17. In (28) let $r = 2$ and $\mathbf{c} = [1, 1, 1, 1]^{\mathrm{T}}$. (Thus, $N = 5$.)

(a) Solve for $\mathbf{U}_{k+1}$ using Jacobi iteration, terminating the iterations when $\left|\mathbf{U}_{j,k+1}^{(n+1)} - \mathbf{U}_{j,k+1}^{(n)}\right| < 0.0001$, say, for each j ($j = 1, 2, 3, 4$). Record the number of iterations needed to achieve that accuracy.

(b) Same as part (a) but using Gauss–Seidel iteration instead.

(c) Same as part (a) but using the SOR method. Use $\omega = 0.9, 1.0, 1.1$, and 1.2, and compare the optimum ω with that predicted by (39).

18. (*Convergence of Jacobi iteration*) It is stated below (33) that the Jacobi iteration (33) converges to the solution of (26) for all finite values of r. Prove that claim. HINT: Denote the eigenvalues and eigenvectors of $\mathbf{A}'$ as $\lambda_1, \dots, \lambda_{N-1}$ and $\mathbf{\Phi}_1, \dots, \mathbf{\Phi}_{N-1}$. Expanding $\mathbf{c}$ as

$$\mathbf{c} = \sum_1^{N-1} c_j \mathbf{\Phi}_j, \tag{18.1}$$

(31) and (33) give

$$\begin{aligned}
\mathbf{U}_{k+1}^{(0)} &= \beta \sum_1^{N-1} c_j \mathbf{\Phi}_j, \qquad \left(\beta \equiv \frac{1}{2(1+r)}\right) \\
\mathbf{U}_{k+1}^{(1)} &= \beta \sum_1^{N-1} c_j \mathbf{\Phi}_j - \beta^2 \sum_1^{N-1} c_j \lambda_j \mathbf{\Phi}_j \\
&= \beta \sum_1^{N-1} (1 - \beta\lambda_j) c_j \mathbf{\Phi}_j, \\
\mathbf{U}_{k+1}^{(2)} &= \beta \sum_1^{N-1} (1 - \beta\lambda_j + \beta^2\lambda_j^2) c_j \mathbf{\Phi}_j,
\end{aligned}$$

and so on. Use (13.11) in Exercise 13 to show that $|\beta\lambda_j| < 1$ for each j, for all $r < \infty$. Thus, show that

$$\mathbf{U}_{k+1}^{(n)} \to \beta \sum_{1}^{N-1} \frac{1}{1+\beta\lambda_j} \mathbf{\Phi}_j \tag{18.2}$$

as $n \to \infty$, and verify that (18.2) satisfies (26).

Chapter 18 Review

The central problem of this chapter is the one-dimensional diffusion equation

$$\alpha^2 u_{xx} = u_t \tag{1}$$

on a finite x interval, with constant Dirichlet (u given) or Neumann (u_x given) boundary conditions. The problem can be solved by the method of separation of variables, the key point of which is the reduction of the PDE to two ODE's governing the factors $X(x)$ and $T(t)$. We emphasize that the boundary conditions are to be applied before the initial condition and we show how to use Fourier series to satisfy the initial condition. Our approach is only formal, but the issue of rigorous justification is addressed in the brief optional Section 18.3.2.

In other cases Fourier series (i.e., the half- and quarter-range formulas and the full Fourier series of a periodic function) may not suffice – for instance, if we have Robin boundary conditions (a linear combination of u and u_x given) or if we have axisymmetric heat conduction in a circular disk or cylinder, governed by the PDE

$$\alpha^2 \left(u_{rr} + \frac{1}{r} u_r \right) = u_t. \tag{2}$$

These cases can be handled using the more powerful Sturm–Liouville theory, as discussed in the optional Section 18.3.3, and the series expansion required for satisfaction of the initial condition is in terms of the eigenfunctions of the relevant Sturm–Liouville problem.

Additional generalizations of (1) are contained in the exercises for Section 18.3, such as the inclusion of one or more of the terms Vu_x (due to axial convection of the medium), Hu (a Newton cooling term due to lateral heat loss), and $F(x,t)$ (due to a distributed source within the medium) in the equation

$$\alpha^2 u_{xx} = u_t + Vu_x + Hu - F(x,t). \tag{3}$$

The Fourier and Laplace transforms enable us to handle problems for which the basic separation of variable method fails or is awkward – for instance, problems on a semi-infinite ($0 < x < \infty$) or infinite ($-\infty < x < \infty$) x domain, or with nonconstant boundary conditions or a distributed source term. These cases are discussed in Section 18.4.

The optional Section 18.5 explains a useful method known as the method of images, which is based on the idea of satisfying homogeneous boundary conditions by fictitiously extending the problem so as to build in symmetry or antisymmetry about the boundary in question. The method is used again in Chapters 19 and 20, but the idea is simple enough so that Section 18.5 is *not* a prerequisite for those chapters.

The final section, 18.6, explains the numerical solution of the diffusion equation by the method of finite differences and gives a glimpse of the power of numerical simulation. We find that the simple explicit method is limited by the condition that

$$r = \alpha^2 \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2} \tag{4}$$

for convergence and stability. To remove the condition (4) we can use an implicit method instead, such as the Crank–Nicolson method, but the price that we pay is that to generate each time line of $U_{j,k}$ values we need to solve an $(N-1) \times (N-1)$ matrix equation, where N is the number of divisions of L (i.e., $\Delta x = L/N$). Fortunately, the matrix is strongly diagonal, so the equation can be solved efficiently by iteration. The mathematics of the finite-difference method is linear algebra since one is replacing a linear PDE by a large system of linear algebraic equations. Questions of convergence, stability, and efficiency are best dealt with using linear algebra methods such as are covered here in Chapters 8–12.

Important general features of the diffusion equation are as follows:

1. The diffusion equation is an *initial-value problem* in t, as is especially clear in Section 18.6 where we see from the "marching" scheme

$$U_{j,k+1} = rU_{j-1,k} + (1-2r)U_{j,k} + rU_{j+1,k} \tag{5}$$

   that the solution along the line $t = (k+1)\Delta t$ is implied by the solution along the preceding line $t = k\Delta t$.

2. Diffusion is a *smoothing process*, as seen from our various solution plots and also from the finite-difference formula (5), since $U_{j,k+1}$ is thereby given as a weighted average of the preceding values $U_{j-1,k}$, $U_{j,k}$ and $U_{j+1,k}$, "average" because the coefficients $r, 1-2r, r$ in (5) sum to unity.

Recall from Section 18.2 that the linear PDE

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f \tag{6}$$

is **parabolic** if $B^2 - AC = 0$, and that the diffusion equation (1) is a simple or canonical form of the general second-order parabolic PDE (with the y's changed to t's). Thus, understand (1) to be *representative* of that entire class of PDE's.

Chapter 19

Wave Equation

19.1 Introduction

The **wave equation**

$$\boxed{c^2\nabla^2 u = u_{tt}} \tag{1}$$

governs a wide variety a wave phenomena such as electromagnetic waves, water waves, supersonic flow, pulsatile blood flow, acoustics, elastic waves in solids, and vibrating strings and membranes. In this introductory section we derive the wave equations governing the vibrating string and vibrating membrane and outline, in the exercises, several other such cases leading to wave equations.

By a *vibrating string* we mean a taut string, such as a guitar string, undergoing a planar vibratory motion. The problem of the vibrating string is of historical importance since it was studied extensively by the great mathematicians *Leonhard Euler* (1707–1783), *Jean Le Rond d'Alembert* (1717–1783), *Daniel Bernoulli* (1700–1782), and *Joseph-Louis Lagrange* (1736–1813). That work gave birth to the subject of partial differential equations. For example, it was in his study of the vibrating string that d'Alembert developed the method of separation of variables used in Chapter 18 and which is now a standard tool in the solution of PDE's. Likewise, he found it necessary to represent the initial shape of the string by what is now known as a Fourier series. Whether such representation is possible, for any given initial shape of the string, was the subject of heated debate, a debate that continued well into the nineteenth century. The theory of Fourier series that emerged is the subject of Chapter 17.

Beginning our derivation, consider a flexible string stretched under tension τ newtons between fixed endpoints at $x = 0$ and $x = L$ on a horizontal x axis (Fig. 1). Let the mass per unit length of the string be σ, a constant, and suppose that the string supports a distributed load $f(x,t)$ newtons per unit x length, counted as positive if it acts downward. Considering the plane vertical motion of the string (in the x, y plane), the desired unknown is the displacement $y(x,t)$.

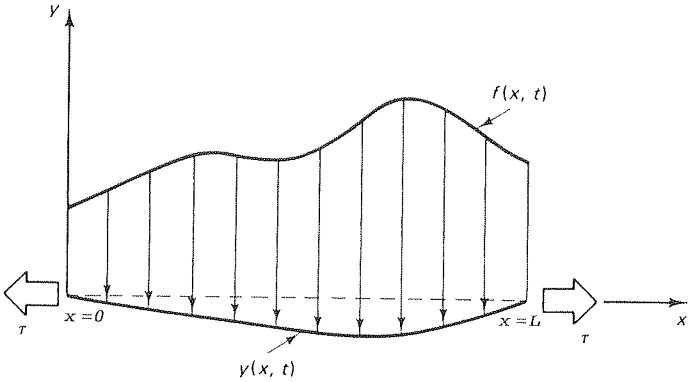


Figure 1. Loaded vibrating string.

We assume that

1. the slope $\partial y/\partial x$ is uniformly small over the length of the string (i.e., $|\partial y/\partial x| \ll 1$);
2. acting on each cross section of the string is the tangentially oriented tension force τ, but no shear force or bending moment as is explained in Example 2 of Section 1.3.

The relevant physical principle is Newton's second law of motion, which we apply to the vertical motion of an element of the string between x and $x + \Delta x$ (Fig. 2):

$$\begin{aligned}\tau \sin\theta(x+\Delta x, t) - \tau\sin\theta(x,t) - f(x+\alpha\Delta x, t)\Delta x \\ = \sigma\Delta s\,\frac{\partial^2 y}{\partial t^2}(x+\beta\Delta x, t),\end{aligned} \tag{2}$$

where Δs is the arc length. That is, the sum of the vertical forces [on the left-hand side of (2)] equals the mass $\sigma\Delta s$ times the vertical acceleration of the mass center at $x + \beta\Delta x$ (for some β such that $0 \le \beta \le 1$). We have assumed that f is a continuous function of x (although that condition could be relaxed) so that there is a point $x + \alpha\Delta x$ (for some α between 0 and 1) at which f takes on its average value over the $(x, x + \Delta x)$ interval.

According to Assumption 1, θ is small so we have these approximations from the Taylor series of $\sin\theta$ and $\tan\theta$:

$$\sin\theta = \theta - \frac{1}{3!}\theta^3 + \frac{1}{5!}\theta^5 - \cdots \approx \theta \tag{3}$$

and

$$\tan\theta = \theta + \frac{1}{3!}\theta^3 + \frac{2}{15}\theta^5 + \cdots \approx \theta. \tag{4}$$

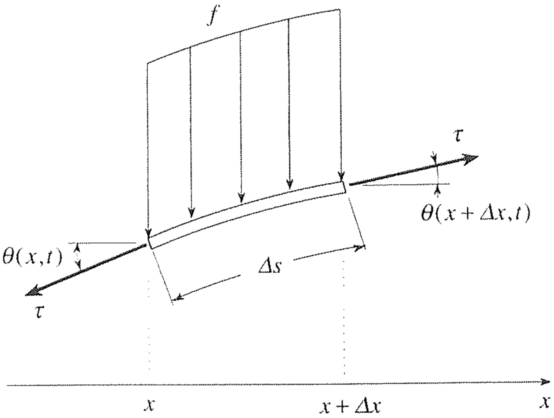


Figure 2. String element.

Hence, for small θ we have $\sin\theta \approx \tan\theta$, the truth of which can also be seen graphically in Fig. 3. But $\tan\theta$ is the slope $\partial y/\partial x$. It also follows from assumption 1 that $\cos\theta = 1 - \frac{1}{2!}\theta^2 + \cdots \approx 1$, so $\Delta s = \Delta x/\cos\theta \approx \Delta x$. Thus, we can eliminate the temporary variables θ and s in favor of y and x and re-express (2) as

$$\tau\,\frac{\frac{\partial y}{\partial x}(x+\Delta x,t) - \frac{\partial y}{\partial x}(x,t)}{\Delta x} - f(x+\alpha\Delta x,t) = \sigma\,\frac{\partial^2 y}{\partial t^2}(x+\beta\Delta x,t). \tag{5}$$

Finally, letting $\Delta x \to 0$ in (5) gives the desired PDE

$$\tau\,\frac{\partial^2 y}{\partial x^2}(x,t) - f(x,t) = \sigma\,\frac{\partial^2 y}{\partial t^2}(x,t) \tag{6}$$

governing $y(x,t)$. If f is simply the gravitational force on the string (i.e., the weight force per unit length of the string), then $f(x,t) = \sigma g =$ constant, where g is the acceleration due to gravity, and (6) becomes

$$\tau\,\frac{\partial^2 y}{\partial x^2} = \sigma\,\frac{\partial^2 y}{\partial t^2} + \sigma g. \tag{7}$$

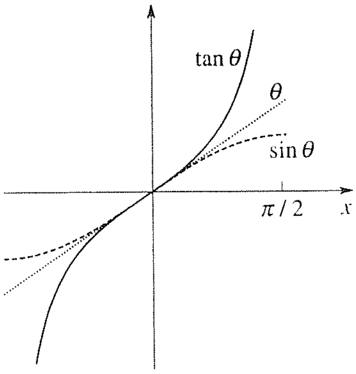


Figure 3. $\sin\theta \approx \theta \approx \tan\theta$ for small θ.

If the gravitational term is negligible, as is surely the case for a guitar string,* then (7) reduces to

$$\tau\,\frac{\partial^2 y}{\partial x^2} = \sigma\,\frac{\partial^2 y}{\partial t^2} \tag{8}$$

or, setting $c \equiv \sqrt{\tau/\sigma}$ and using the more compact subscript notation for partial derivatives,

$$\boxed{c^2 y_{xx} = y_{tt},} \tag{9}$$

*A guitar sounds the same if it is played horizontally [so the σg term is present in (7)] or vertically [so the σg term is *not* present in (7)]. Applying dimensional reasoning to (7), we find that the σg term can be neglected if $\sigma g << \tau/L$.

which is the classical *one-dimensional wave equation* governing $y(x, t)$.

It is always important to tie together the mathematics and the physics. Accordingly, how are we to understand the terms in (8)? Essentially, (8) is of the form force = mass × acceleration.* The τy_{xx} term is the net vertical force on the element due to the tension τ. To understand its form, recall from the calculus that the local radius of curvature R and curvature κ for a plane curve $y(x)$ (Fig. 4) are given by

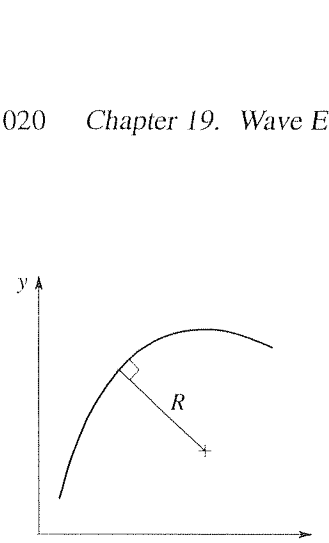


Figure 4. Radius of curvature R.

$$\kappa = \frac{1}{R} = \frac{y''}{(1+y'^2)^{3/2}}. \tag{10}$$

In our case $|y'| \ll 1$, so (10) becomes $\kappa \approx y''$. Thus, within our assumption of small deflection and small slope the τy_{xx} term in (8) is the product of the tension and the curvature. That result makes sense physically because it is through the curvature that the two tension forces in Fig. 2 are misaligned and therefore have a nonzero vertical resultant.

We close this section with an introduction to the *two*-dimensional wave equation. Specifically, consider the two-dimensional version of a vibrating string, a vibrating *membrane* such as a drumhead. We assume that the membrane is stretched uniformly under a tension τ per unit length. That is, at each point of the membrane the tension per unit length along any straight line through that point, independent of the orientation of the line, is τ. If, for example, we stretch a rectangular membrane horizontally (Fig. 5) and then clamp the four edges, then the membrane would *not* be stretched uniformly, for the tension per unit length along any vertical line would be τ whereas along any horizontal line it would be zero.

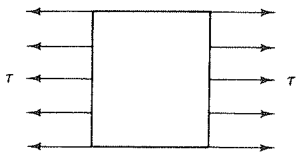


Figure 5. *Non*uniformly stretched membrane.

Denoting the displacement of the membrane out of the x, y plane as $w(x, y, t)$, we proceed essentially as before. Thus, we assume that the slopes $\partial w/\partial x$ and $\partial w/\partial y$ are uniformly small over the domain (i.e., $|\partial w/\partial x| \ll 1$ and $|\partial w/\partial y| \ll 1$), and that the membrane is perfectly flexible, so that only the tangential tensile force τ acts. Then, applying Newton's second law to a membrane element lying between x and $x + \Delta x$ and between y and $y + \Delta y$ (Fig. 6), gives

$$\begin{aligned} &\tau \Delta y \sin\theta\Big|_{x+\Delta x} - \tau \Delta y \sin\theta\Big|_{x} \\ &+\tau \Delta x \sin\phi\Big|_{y+\Delta y} - \tau \Delta x \sin\phi\Big|_{y} - f \Delta x \Delta y = \sigma \Delta A \, \frac{\partial^2 w}{\partial t^2}, \end{aligned} \tag{11}$$

where σ is the mass per unit area of the membrane, ΔA is the surface area of the element under consideration, and $f(x, y, t)$ is a distributed load counted as positive if it acts downward. By virtue of the stated assumptions, $\sin\theta \approx \theta \approx \tan\theta = \partial w/\partial x$, $\sin\phi \approx \phi \approx \tan\phi = \partial w/\partial y$, and $\Delta A \approx \Delta x \Delta y$, so (11) becomes

$$\tau \, \frac{w_x\big|_{x+\Delta x} - w_x\big|_x}{\Delta x} + \tau \, \frac{w_y\big|_{y+\Delta y} - w_y\big|_y}{\Delta y} - f = \sigma w_{tt}, \tag{12}$$

*More precisely, (8) is on a per unit x-length basis: the vertical force per unit x length is equal to the mass per unit x length times the vertical acceleration.

and letting $\Delta x \to 0$ and $\Delta y \to 0$ gives the PDE

$$\tau(w_{xx} + w_{yy}) - f(x, y, t) = \sigma w_{tt}. \tag{13}$$

If $f = 0$ and if we define $c \equiv \sqrt{\tau/\sigma}$, then (13) becomes the classical *two-dimensional wave equation*

$$\boxed{c^2(w_{xx} + w_{yy}) = w_{tt}} \tag{14}$$

governing $w(x, y, t)$.

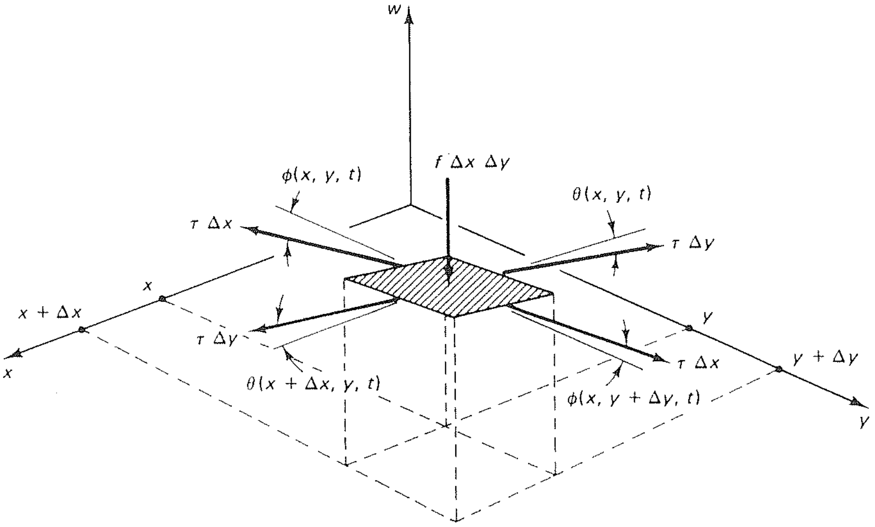


Figure 6. Membrane element.

In the next two sections we solve the vibrating string and vibrating membrane equations by separation of variables.

Closure. In this section we derive the one- and two-dimensional wave equations (9) and (14) governing vibrating strings and vibrating membranes, respectively, subject to the assumptions that the string or membrane is flexible and that the slopes are small. Several additional problems governed by wave equations are given in the exercises.

EXERCISES 19.1

1. (*Hanging chain*) Let a flexible chain hang from the ceiling. Measure x downward from the ceiling and let $y(x, t)$ be its lateral displacement. Modifying our derivation of (9) as appropriate, show that the PDE governing $y(x, t)$ is

$$g[(L - x)y_x]_x = y_{tt}, \tag{1.1}$$

where g is the acceleration due to gravity and L is the length of the chain.

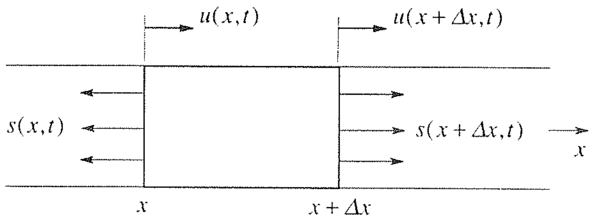
2. (*Longitudinal waves in a rod*) Consider a uniform metal bar, of cross-sectional area A and mass per unit length σ, with a stress distribution $s(x,t)$ and a resulting longitudinal displacement $u(x,t)$. Applying Newton's second law to an element of the rod between x and $x+\Delta x$, shown in the figure, show that

$$A\frac{\partial s}{\partial x} = \sigma\frac{\partial^2 u}{\partial t^2}. \tag{2.1}$$

Suppose that the material admits a linear stress-strain relationship $s = E\epsilon$, where the constant of proportionality E is Young's modulus, and the strain ϵ is defined as the "stretch per unit length," so that $\epsilon = \dfrac{u(x+\Delta x,t) - x(x,t)}{\Delta x}$. From these relations, show that s and u both satisfy one-dimensional wave equations

$$c^2 u_{xx} = u_{tt} \qquad \text{and} \qquad c^2 s_{xx} = s_{tt}, \tag{2.2}$$

where $c \equiv \sqrt{EA/\sigma}$.



3. (*Electromagnetic waves*) (a) Electromagnetic fields in free space (i.e., in a vacuum) are governed by the famous **Maxwell's equations**:

$$\begin{aligned}
\nabla\times\mathbf{H} &= \epsilon_0\frac{\partial\mathbf{E}}{\partial t},\\
\nabla\times\mathbf{E} &= -\mu_0\frac{\partial\mathbf{H}}{\partial t},\\
\nabla\cdot\mathbf{H} &= 0,\\
\nabla\cdot\mathbf{E} &= 0,
\end{aligned} \tag{3.1,2,3,4}$$

where $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field intensities, respectively, and where ϵ_0 and μ_0 are the permittivity and permeability of free space, respectively. Show that it follows from (3.1)–(3.4) that $\mathbf{E}$ and $\mathbf{H}$ (which are functions of x, y, z and the time t) satisfy the wave equations

$$c^2\nabla^2\mathbf{H} = \mathbf{H}_{tt} \qquad \text{and} \qquad c^2\nabla^2\mathbf{E} = \mathbf{E}_{tt}, \tag{3.5}$$

where $c \equiv 1/\sqrt{\epsilon_0\mu_0}$. HINT: Recall the identity $\nabla\times(\nabla\times\mathbf{v}) = \nabla(\nabla\cdot\mathbf{v}) - \nabla^2\mathbf{v}$ from Section 16.6.

(b) Write the x, y, z components of the vector wave equations (3.5).

4. (*Current and magnetic field*) Suppose that the "string" is actually a flexible wire of mass per unit length σ, under tension τ, carrying a current I in the presence of a uniform magnetic field of magnetic flux density $\mathbf{B} = B_1\hat{\mathbf{i}} + B_2\hat{\mathbf{j}} + B_3\hat{\mathbf{k}}$. Since a charge Q moving with velocity $\mathbf{U}$ in such a field experiences a force $\mathbf{F} = Q\mathbf{U}\times\mathbf{B}$, we may need to revise the vibrating string equation to include lateral magnetic forcing terms.

(a) Show that within the usual vibrating string approximations, the magnetic force exerted on an element of the wire is

$$\Delta\mathbf{F} \approx qA\Delta x\left[U\hat{\mathbf{i}} + (Uy_x + y_t)\hat{\mathbf{j}} + (Uz_x + z_t)\hat{\mathbf{k}}\right]\times\mathbf{B}, \tag{4.1}$$

where the constants q, A, U are the charge per unit volume within the wire, cross-sectional area of the wire, and velocity of the charges within the wire, respectively. HINT: The x, y, z velocity components of an element of charge within the wire are

$$\begin{aligned}
\frac{dx}{dt} &= U,\\
\frac{dy}{dt}(x(t),t) &= y_x\frac{dx}{dt} + y_t = Uy_x + y_t,\\
\frac{dz}{dt}(x(t),t) &= z_x\frac{dx}{dt} + z_t = Uz_x + z_t.
\end{aligned}$$

(b) Working out the cross product in (4.1) and noting that qAU is the current I, show that the equations of (the not necessarily planar) motion of the wire are

$$\begin{aligned}
\tau y_{xx} - IB_3 + (Iz_x + qAz_t)B_1 &= \sigma y_{tt},\\
\tau z_{xx} + IB_2 - (Iy_x + qAy_t)B_1 &= \sigma z_{tt},
\end{aligned} \tag{4.2}$$

where x, y, z are right-handed Cartesian coordinates. Notice that these equations are *coupled* due to the z_x, z_t, y_x, and y_t terms. Naturally, in a particular application one or more of the additional magnetic-field terms might be negligible.

5. (*Water waves*) Consider plane water waves in water of depth $h(x)$ as shown in the figure. If the wavelength is much greater than h (as is true for ocean tides and certain waves in shallow water), the governing equations are found to be

$$\begin{aligned} u_t + uu_x &= -g\eta_x, \\ [u(\eta + h)]_x &= -\eta_t, \end{aligned} \tag{5.1,2}$$

where u is the x velocity [which is approximately constant with respect to y so $u = u(x,t)$], $\eta(x,t)$ is the free-surface elevation relative to the undisturbed water level, and g is the acceleration due to gravity. If we restrict our attention to small amplitude motions – so that the "second-order" term uu_x can be neglected relative to the "first-order" terms u_t and $g\eta_x$, and η can be neglected relative to h – then show that η satisfies the equation $g(h\eta_x)_x = \eta_{tt}$ or, if $h(x)$ is a constant,

$$c^2\eta_{xx} = \eta_{tt}, \tag{5.3}$$

where $c \equiv \sqrt{gh}$.

19.2 Separation of Variables; Vibrating String

19.2.1. Solution by separation of variables. In Section 19.1 we derive the wave equation governing the motion of a vibrating string. In the present section we complete the formulation by appending boundary and initial conditions and then show how to solve for $y(x,t)$ by the method of separation of variables.

Specifically, let us consider a finite string extending over $0 < x < L$, tied at its ends, and having initial displacement $y(x,0) = f(x)$ and initial velocity $y_t(x,0) = g(x)$, where f and g are prescribed. Thus, the complete problem statement is

$$\begin{aligned} c^2 y_{xx} &= y_{tt}, \qquad (0 < x < L,\ 0 < t < \infty) && \text{(1a)} \\ y(0,t) &= 0,\ \ y(L,t) = 0, \qquad (0 < t < \infty) && \text{(1b)} \\ y(x,0) &= f(x),\ \ y_t(x,0) = g(x) \qquad (0 < x < L) && \text{(1c)} \end{aligned}$$

and is summarized in Fig. 1.

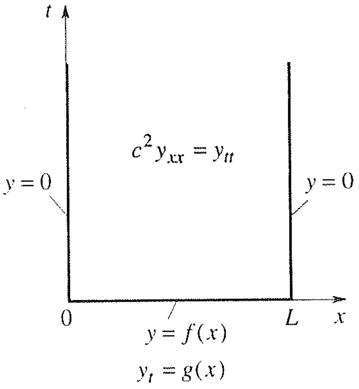


Figure 1. Problem (1) in the x,t plane.

Notice carefully that for the diffusion equation $\alpha^2 u_{xx} = u_t$ (Chapter 18) we prescribe only u initially, whereas for the wave equation we prescribe both y *and* y_t initially. Intuitively, it certainly seems reasonable that to predict $y(x,t)$ we will need to know how the string is set in motion, namely, both the initial displacement $y(x,0)$ and the initial velocity $y_t(x,0)$. From a mathematical point of view we are guided by the fact that the wave equation (1a) is a second-order equation with respect to t (whereas the diffusion equation is of first order), so we expect to need two initial conditions. Verification that the problem statement (1) *uniquely* determines $y(x,t)$ is left for the exercises. Here, we limit our attention to *finding* the solution.

Using separation of variables, seek

$$y(x,t) = X(x)T(t). \tag{2}$$

Putting (2) into (1a) and separating the variables gives

$$\frac{X''}{X} = \frac{1}{c^2}\frac{T''}{T}. \tag{3}$$

Since the left-hand side of (3) is a function of x alone and the right-hand side is a function of t alone, it follows from (3) (as discussed in Section 18.3) that both sides are constants, say $-\kappa^2$, so

$$\frac{X''}{X} = \frac{1}{c^2}\frac{T''}{T} = \text{constant} = -\kappa^2. \tag{4}$$

Thus,

$$X'' + \kappa^2 X = 0, \tag{5a}$$
$$T'' + \kappa^2 c^2 T = 0, \tag{5b}$$

and

$$X = \begin{cases} A + Bx, & \kappa = 0 \\ D\cos\kappa x + E\sin\kappa x, & \kappa \neq 0 \end{cases} \tag{6}$$

$$T = \begin{cases} H + It, & \kappa = 0 \\ J\cos\kappa ct + K\sin\kappa ct, & \kappa \neq 0. \end{cases} \tag{7}$$

Our motivation for writing the minus sign in (4) is that the resulting ODE (5b) admits $\cos\kappa ct$ and $\sin\kappa ct$ solutions, which look correct since we anticipate a vibratory motion; if we write κ^2 in (4) instead of $-\kappa^2$, then we obtain $T'' - \kappa^2 c^2 T = 0$, instead of (5b), with nonvibratory exponential solutions.*

Thus, we have found product solutions of the form $(A + Bx)(H + It)$ and $(D\cos\kappa x + E\sin\kappa x)(J\cos\kappa ct + K\sin\kappa ct)$ and, relying on the linearity of $L[y] = c^2 y_{xx} - y_{tt} = 0$, we can use superposition and write

$$\begin{aligned} y(x,t) = {} & (A + Bx)(H + It) \\ & + (D\cos\kappa x + E\sin\kappa x)(J\cos\kappa ct + K\sin\kappa ct). \end{aligned} \tag{8}$$

As in Chapter 18, we apply the boundary conditions before the initial conditions. Accordingly,

$$y(0,t) = 0 = A(H + It) + D(J\cos\kappa ct + K\sin\kappa ct). \tag{9}$$

Since the right-hand side of (9) is a linear combination of the linearly independent functions $1, t, \cos\kappa ct, \sin\kappa ct$, it follows from (9) that we must have either $A = 0$ or $H = I = 0$, and either $D = 0$ or $J = K = 0$. We choose $A = 0$ and $D = 0$ so as to be left with as robust a solution as possible, namely,

$$y(x,t) = Bx(H + It) + E\sin\kappa x(J\cos\kappa ct + K\sin\kappa ct) \tag{10}$$

or, combining BH as P, BI as Q, EJ as R, and EK as S, for brevity,

$$y(x,t) = x(P + Qt) + \sin\kappa x(R\cos\kappa ct + S\sin\kappa ct). \tag{11}$$

*As discussed in Exercise 2 of Section 18.3, we can survive writing κ^2 instead of $-\kappa^2$, but the choice is less convenient because κ will then end up being purely imaginary rather than real.

In case it is not clear how the choice $A = D = 0$ gives "as robust a solution as possible," let us return to (9) and focus on the $D(J\cos\kappa ct + K\sin\kappa ct)$ term. If we choose $J = K = 0$, then we lose the entire $(D\cos\kappa x + E\sin\kappa x)(J\cos\kappa ct + K\sin\kappa ct)$ term in (8), whereas if we choose $D = 0$ as, indeed, we did, then we are left with $E\sin\kappa x(J\cos\kappa ct + K\sin\kappa ct)$. Similarly, if we infer from (9) that $H = I = 0$, then we lose the entire $(A + Bx)(H + It)$ term in (8), whereas if we infer that $A = 0$, then we are left with $Bx(H + It)$.

Next,

$$y(L,t) = 0 = L(P + Qt) + \sin\kappa L(R\cos\kappa ct + S\sin\kappa ct) \tag{12}$$

so we need

$$LP = 0 \quad \text{and} \quad LQ = 0 \tag{13}$$

as well as

$$R\sin\kappa L = 0 \quad \text{and} \quad S\sin\kappa L = 0. \tag{14}$$

Since $L \neq 0$, (13) gives $P = Q = 0$. And if we are to avoid having $R = S = 0$, we must choose

$$\sin\kappa L = 0. \tag{15}$$

Thus, $\kappa = n\pi/L$ for $n = 1, 2, \ldots$. Putting these results into (11) and using superposition gives

$$y(x,t) = \sum_{n=1}^{\infty} \sin\frac{n\pi x}{L}\left(R_n\cos\frac{n\pi ct}{L} + S_n\sin\frac{n\pi ct}{L}\right). \tag{16}$$

[If any steps in deriving (16) are unclear, we urge you to review Example 1 in Section 18.3]

Our expectation is that the R's and S's can now be determined from the initial conditions (1c). Imposing those conditions gives

$$y(x,0) = f(x) = \sum_{n=1}^{\infty} R_n\sin\frac{n\pi x}{L} \qquad (0 < x < L) \tag{17a}$$

and

$$y_t(x,0) = g(x) = \sum_{n=1}^{\infty} \frac{n\pi c}{L}S_n\sin\frac{n\pi x}{L}. \qquad (0 < x < L) \tag{17b}$$

We can identify each of (17a,b) as a half-range sine series so (→ pág 870)

$$R_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx \tag{18a}$$

and

$$\frac{n\pi c}{L}S_n = \frac{2}{L}\int_0^L g(x)\sin\frac{n\pi x}{L}\,dx$$

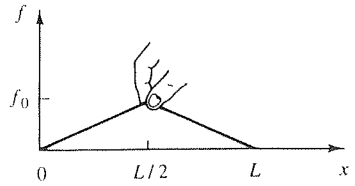


Figure 2. Plucked string.

or

$$S_n = \frac{2}{n\pi c}\int_0^L g(x)\sin\frac{n\pi x}{L}\,dx. \tag{18b}$$

Hence, the solution of (1) is given by (16) with the R_n's and S_n's computed according to (18a) and (18b).

To illustrate, let $f(x)$ be as shown in Fig. 2 and let $g(x) = 0$. That is, we pull the string up at its midpoint and then release it from rest. Then (18a,b) give

$$R_n = \frac{8f_0}{n^2\pi^2}\sin\frac{n\pi}{2} \quad \text{and} \quad S_n = 0, \tag{19}$$

so

$$y(x,t) = \frac{8f_0}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}\sin\frac{n\pi}{2}\sin\frac{n\pi x}{L}\cos\frac{n\pi ct}{L}. \tag{20}$$

The right-hand side of (20) is a superposition of distinct **modes** of vibration $\sin(n\pi x/L)\cos(n\pi ct/L)$, each of which is a **standing wave** of spatial frequency $n\pi/L$ and temporal frequency $n\pi c/L$. The first three mode shapes are depicted in Fig. 3. The points $x = 0$ and $x = L$ (heavy dots in Fig. 3) are called **nodal points** of the first mode because $y = 0$ there for all t; $x = 0$, $L/2$, and L are nodal points of the second mode, and so on. Further, we say that the modes are **orthogonal** inasmuch as their shapes satisfy the **orthogonality relation**

$$\int_0^L \sin\frac{m\pi x}{L}\sin\frac{n\pi x}{L}\,dx = 0 \tag{21}$$

for any pair of integers m and n with $m \neq n$ [equation (24b), Section 17.3].

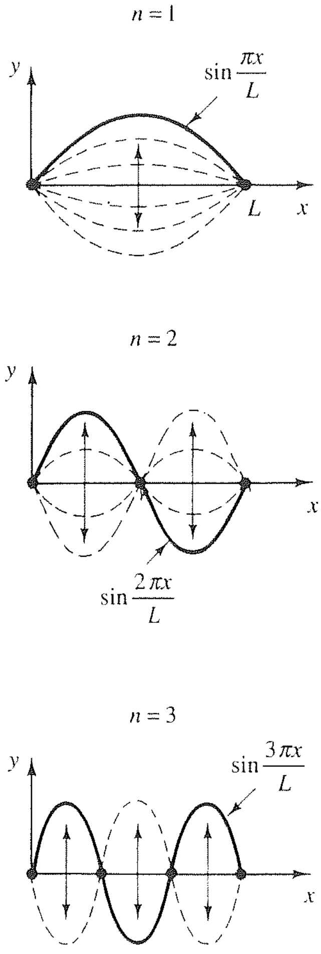


Figure 3. The first three modes.

It is interesting to see what (20) can tell us about the musical quality of a violin string plucked in the manner shown in Fig. 2. Suppose, for definiteness, that we tune the string so that its fundamental frequency $\pi c/L$ corresponds to the lowest A on a piano, say A_0. Since A_0's frequency is 27.5 cycles/sec, we accomplish that tuning by adjusting the tension τ so that

$$\frac{\pi c}{L} = \frac{\pi}{L}\sqrt{\frac{\tau}{\sigma}}\,\frac{\text{rad}}{\text{sec}} = \left(27.5\,\frac{\text{cycles}}{\text{sec}}\right)\left(\frac{2\pi\,\text{rad}}{\text{cycle}}\right), \tag{22}$$

or

$$\tau = (55L)^2\sigma. \tag{23}$$

Then the first several terms in (20) correspond to the combination of notes shown in Table 1, where A_1 is an A one octave higher than A_0 , and so on, and $C_3^{\#}$ is C sharp in the third octave above the lowest. Observe that the overtones (with nonzero amplitude), $E_2, C_3^{\#}, \ldots$, do not occur at octaves above A_0. Thus, the sound is not a pristine A_0 with octave overtones but it is, nonetheless, fairly "clean" because of the relatively small amplitudes of the $E_2, C_3^{\#}, \ldots$ contributions. The mix of frequencies and amplitudes is different for different instruments and an A_0 played on

Table 1. The first five notes in (20).

n	Frequency $n\dfrac{\pi c}{L}$ (cycles/sec)	Relative Amplitude $\dfrac{1}{n^2}\left\|\sin\dfrac{n\pi}{2}\right\|$	Musical Note
1	27.5	1	A_0
2	55.0	0	A_1
3	82.5	$\frac{1}{9}$	$\approx E_2$
4	110.0	0	A_2
5	137.5	$\frac{1}{25}$	$\approx C_3^{\#}$

a violin sounds different from an A_0 played on a tuba.*

19.2.2. Traveling wave interpretation. We have seen that if $y(x,0) = f(x)$ is prescribed and $y_t(x,0) = 0$, then $y(x,t)$ is given by

$$y(x,t) = \sum_{n=1}^{\infty} R_n \sin\frac{n\pi x}{L}\cos\frac{n\pi ct}{L}, \qquad (0 < x < L,\ 0 < t < \infty) \tag{24}$$

where the R_n's are computed from the initial condition

$$y(x,0) = f(x) = \sum_{n=1}^{\infty} R_n \sin\frac{n\pi x}{L} \qquad (0 < x < L) \tag{25}$$

as

$$R_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx. \tag{26}$$

If we wish to plot $y(x,t)$ at various times, we can sum the series (24) at a number of different x's and t's. However, we can do much better as we will now show. First, use the trigonometric identity

$$\sin A\cos B = \frac{1}{2}\left[\sin(A-B) + \sin(A+B)\right] \tag{27}$$

to re-express (24) as

$$y(x,t) = \frac{1}{2}\sum_{n=1}^{\infty} R_n\left[\sin\frac{n\pi}{L}(x-ct) + \sin\frac{n\pi}{L}(x+ct)\right]$$

*Actually, the sound of a violin is due very little to the air being set in motion by the vibrating string. Rather,the string drives the sounding board, through its connection at the bridge, and it is the vibrating sounding board that sets the adjacent air in motion and creates an audible sound. Thus, a serious investigation of violin mechanics would lead immediately to a much more difficult analysis of the vibrating sounding board.

$$= \frac{1}{2}\left[\sum_{n=1}^{\infty} R_n \sin\frac{n\pi}{L}(x-ct) + \sum_{n=1}^{\infty} R_n \sin\frac{n\pi}{L}(x+ct)\right]. \qquad (28)$$

Finally, comparing (28) with (25) we see that the two series in (28) *can be summed into closed form* in terms of f as

$$y(x,t) = \frac{1}{2}[f(x-ct) + f(x+ct)]. \qquad (29)$$

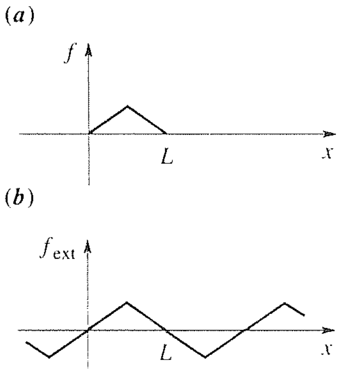


Figure 4. f and $f_{\text{ext}}(x)$.

There is a difficulty associated with (29): for a given x in the interval $(0, L)$ and a given t in the t interval $(0, \infty)$, one or both of the arguments $x - ct$ and $x + ct$ in (29) may lie outside the domain of definition of f, $0 \le x \le L$, in which event (29) is meaningless. However, recall from our study of half-range sine series (Section 17.4) that the half-range sine expansion of $f(x)$ on $0 < x < L$, given by (25), is also the Fourier series expansion of the extended function $f_{\text{ext}}(x)$ that is $2L$-periodic and antisymmetric about $x = 0$ and $x = L$. For example, if $f(x)$ is the function shown in Fig. 4a, then $f_{\text{ext}}(x)$ is the function shown in Fig. 4b. Thus, the right-hand side of (28) actually sums to

$$\boxed{y(x,t) = \frac{1}{2}[f_{\text{ext}}(x-ct) + f_{\text{ext}}(x+ct)]} \qquad (30)$$

for *any* values of the arguments $x - ct$ and $x + ct$.

To illustrate the use of (30), let $L = 10$ and $c = 12$, let f be as shown in Fig. 2 with $f_0 = 1$, and let us compute y at $x = 2$ and $t = 3$. Using (30) and the fact that f_{ext} is periodic with period $2L = 20$ and odd, we have

$$\begin{aligned}
y(2,3) &= \tfrac{1}{2}[f_{\text{ext}}(2-36) + f_{\text{ext}}(2+36)] \\
&= \tfrac{1}{2}[f_{\text{ext}}(-34) + f_{\text{ext}}(38)] \\
&= \tfrac{1}{2}[f_{\text{ext}}(-14) + f_{\text{ext}}(18)] && \text{(periodicity of } f_{\text{ext}}) \\
&= \tfrac{1}{2}[f_{\text{ext}}(6) + f_{\text{ext}}(-2)] && \text{(periodicity of } f_{\text{ext}}) \\
&= \tfrac{1}{2}[f(6) - f(2)] && [f_{\text{ext}} \text{ is odd, and } f_{\text{ext}}(x) = f(x) \\
& && \quad \text{on } 0 < x < 10] \\
&= \tfrac{1}{2}(\tfrac{4}{5} - \tfrac{2}{5}) && \text{(definition of } F \text{ in Fig. 2)} \\
&= \tfrac{1}{5}.
\end{aligned}$$

Besides using (30) to compute y at any specific x and t, we can use it to obtain the graph of $y(x,t)$ over the entire interval $0 < x < L$, at any given t, by observing that the graph of $f_{\text{ext}}(x - ct)$ plotted as a function of x is simply the graph of $f_{\text{ext}}(x)$ translated to the right through a distance ct. Similarly, the graph of $f_{\text{ext}}(x + ct)$ is the graph of $f_{\text{ext}}(x)$ translated to the left through a distance ct. Thus, if f is as shown in Fig. 2, then $f_{\text{ext}}(x)$ is as shown in Fig. 5a, $f_{\text{ext}}(x-ct)$ and $f_{\text{ext}}(x + ct)$ are as shown in Fig. 5b, and $y(x,t) = \frac{1}{2}[f_{\text{ext}}(x - ct) + f_{\text{ext}}(x+ct)]$ is (by addition of the two graphs in Fig. 5b and scaling by $1/2$) as shown in Fig. 5c.

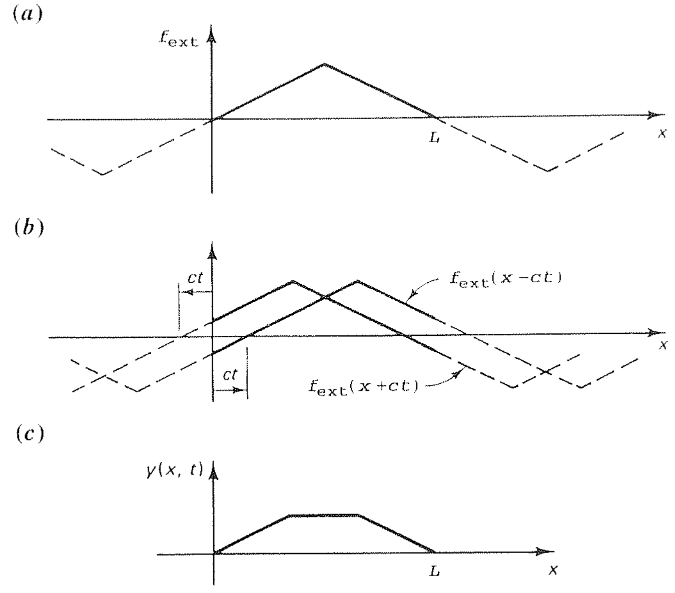


Figure 5. Graphical use of (30).

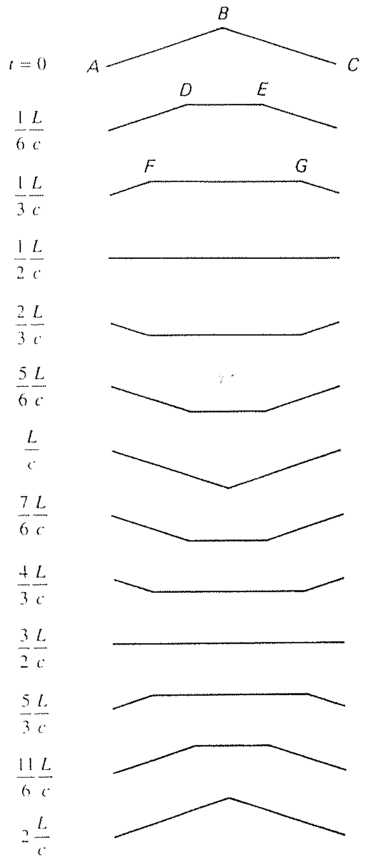


Figure 6. Solution sequence over one cycle.

Carrying out this graphical procedure at a number of different times over one complete cycle yields the solution sequence shown in Fig. 6. Consider the results shown in Fig. 6 in terms of the physics. At $t = 0$ the string segments AB and BC are straight, so each string element is in static equilibrium – except at B, which point is driven downward (Fig. 7). The plateau DE (Fig. 6) moves downward at constant velocity (i.e., with no acceleration) since there is no net vertical force on the elements between D and E. (Remember that we have neglected the effects of gravity.)

We can now explain why $c^2 y_{xx} = y_{tt}$ is called the "wave" equation for we see from (24) that $y(x,t)$ can be expressed as a superposition of **standing waves** of shape $\sin(n\pi x/L)$ and temporal frequency $n\pi c/L$ rad/sec. The sum of these waves is also a standing wave, of temporal frequency $\pi c/L$. Alternatively, (28) expresses $y(x,t)$ as a superposition of **traveling waves** traveling to the right and left with speed c. Thus, the parameter c in the wave equation $c^2 y_{xx} = y_{tt}$ is now seen to be the **wave speed** – that is, the speed of propagation of the traveling waves. Recalling that $c = \sqrt{\tau/\sigma}$, it does seem reasonable physically that the wave speed should increase with the tension τ and decrease with the lineal mass density σ. To reiterate, (28) expresses y as a superposition of sinusoidal traveling waves criss-crossing leftward and rightward with speed c. Over the physical interval $(0, L)$ these waves sum to a standing wave with nodes at $x = 0$ and $x = L$.

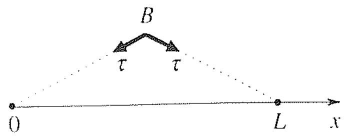


Figure 7. Downward force at B.

19.2.3. Using Sturm–Liouville theory. (Optional) If we rely on the Sturm–Liouville theory, then we understand (17a) and (17b) as eigenfunction expansions of f and g in terms of the orthogonal eigenfunctions $\sin(n\pi x/L)$ of the relevant

Sturm–Liouville problem, namely,

$$X'' + \kappa^2 X = 0, \qquad (0 < x < L) \tag{31a}$$

$$X(0) = 0, \;\; X(L) = 0. \tag{31b}$$

Then the weight function in the inner product is 1, and

$$\begin{aligned} R_n &= \frac{\langle F(x), \sin \frac{n\pi x}{L} \rangle}{\langle \sin \frac{n\pi x}{L}, \sin \frac{n\pi x}{L} \rangle} = \frac{\int_0^L F(x) \sin \frac{n\pi x}{L}\, dx}{\int_0^L \sin^2 \frac{n\pi x}{L}\, dx} \\ &= \frac{2}{L} \int_0^L F(x) \sin \frac{n\pi x}{L}\, dx \end{aligned} \tag{32}$$

as in (18a), and similarly for S_n.

Closure. Like the problem of heat conduction in a finite rod (Section 18.3), the vibrating finite string problem (1) is defined on a semi-infinite strip in the x, t plane, with boundary conditions at $x = 0$ and $x = L$. However, the PDE is of second order with respect to t, so two initial conditions are appropriate rather than one. Solution by separation of variables proceeds in essentially the same manner as in Section 18.3 and, as for the heat equation, it is necessary to apply the boundary conditions before the initial conditions. The solution (24) is in the form of a superposition of orthogonal modes, each one being a standing wave. Alternatively, the trigonometric identity (27) enables us to re-express the solution in the form (28), namely, as a superposition of left- and right-running traveling waves with wave speed c. More simply, (30) gives the solution as the sum of *two* traveling waves, one left-running and one right-running.

From Fig. 6 we see that the initially "kinky" deflection $y(x, t)$ does not smooth out as t increases, the way an initially kinky or discontinuous temperature distribution $u(x, t)$ does, and this is a major difference between the wave and diffusion processes. Rather, kinks and discontinuities in the initial conditions propagate into the x, t domain.

EXERCISES 19.2

1. Let $L = 10$, $c = 12$, and $f_0 = 1$ in (20). Use (20) to compute y at the specified values of x and t, to two significant figures. Then use (30) to compute y and show that your two results agree.

(a) $y(5, 1)$ (b) $y(5, 2)$ (c) $y(5, 3)$
(d) $y(5, 4)$ (e) $y(5, 6)$ (f) $y(5, 10)$
(g) $y(5, 20)$ (h) $y(3, 1)$ (i) $y(3, 10)$
(j) $y(3, 40)$ (k) $y(1, 2)$ (l) $y(1, 10)$

2. Solve (1) for $y(x, t)$ for the case where $f(x) = 0$ and
(a) $g(x) = 50 \sin(\pi x/L)$
(b) $g(x) = 3 \sin(\pi x/L) - 5 \sin(4\pi x/L)$
(c) $g(x) = \sin(2\pi x/L) + \sin(3\pi x/L) + 4 \sin(8\pi x/L)$

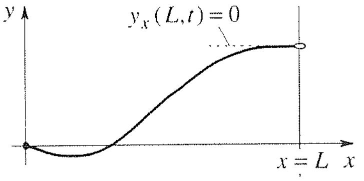
3. Using the solution technique illustrated in Fig. 5, obtain the graphs shown in Fig. 6, for the given values of t, and label any key values.

(a) $t = \dfrac{1}{6}\dfrac{L}{c}$	(b) $t = \dfrac{1}{3}\dfrac{L}{c}$	(c) $t = \dfrac{1}{2}\dfrac{L}{c}$
(d) $t = \dfrac{2}{3}\dfrac{L}{c}$	(e) $t = \dfrac{5}{6}\dfrac{L}{c}$	(f) $t = \dfrac{L}{c}$
(g) $t = \dfrac{7}{6}\dfrac{L}{c}$	(h) $t = \dfrac{4}{3}\dfrac{L}{c}$	(i) $t = \dfrac{3}{2}\dfrac{L}{c}$
(j) $t = \dfrac{5}{3}\dfrac{L}{c}$	(k) $t = \dfrac{11}{2}\dfrac{L}{c}$	(l) $t = 2\dfrac{L}{c}$

4. Construct neat labeled sketches of the graph of $y(x,t)$ versus t for the oscillation depicted in Fig. 6, for $x = L/2$ and also for $x = L/4$.

5. If, instead of the string being tied at $x = L$ such that $y(L,t) = 0$, the string is looped around a vertical frictionless wire (as shown in the figure), then in place of $y(L,t) = 0$ the boundary condition becomes

$$y_x(L,t) = 0. \tag{5.1}$$



(a) Explain why (5.1) is true.
(b) Solve (1a) for $y(x,t)$ by separation of variables, with these boundary and initial conditions:

$$y(0,t) = 0, \quad y_x(L,t) = 0,$$
$$y(x,0) = f(x), \quad y_t(x,0) = 0,$$

leaving expansion coefficients in integral form.
(c) Solve (1a) for $y(x,t)$ by separation of variables, with these boundary and initial conditions:

$$y_x(0,t) = 0, \quad y_x(L,t) = 0,$$
$$y(x,0) = 0, \quad y_t(x,0) = V,$$

where V is a constant.
(d) Solve (1a) for $y(x,t)$ by separation of variables, with these boundary and initial conditions.

$$y_x(0,t) = 0, \quad y(L,t) = 0,$$
$$y(x,0) = 0, \quad y_t(x,0) = g(x),$$

leaving expansion coefficients in integral form.

6. (*Inclusion of damping*) Unless the string vibrates in a vacuum, there will be some damping due to the movement of the string through the fluid (be it air, water, or whatever). If the damping force is proportional to the velocity y_t, the modified equation of motion becomes

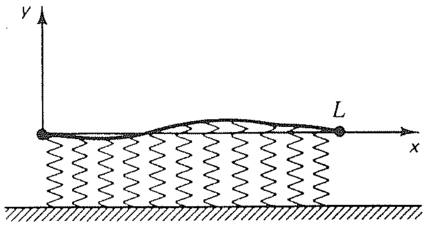
$$c^2 y_{xx} = y_{tt} + a y_t, \tag{6.1}$$

where a is a known constant. For definiteness, suppose that $0 < a < 2\pi c/L$. Solve (6.1) by separation of variables, subject to the conditions.

$$y(0,t) = 0, \quad y(L,t) = 0,$$
$$y(x,0) = f(x), \quad y_t(x,0) = 0,$$

leaving expansion coefficients in integral form. Summarize, in words, the effect(s) of the damping term ay_t.

7. (*Inclusion of lateral spring*) If, as shown in the figure, a lateral distributed spring is included, then the modified equation of motion for the vibrating string is $\tau y_{xx} - ky = \sigma y_{tt}$, where k is the spring stiffness per unit length (newtons per meter per meter) or



$$c^2 y_{xx} - by = y_{tt}. \qquad \left(c^2 = \frac{\tau}{\sigma}, \; b = \frac{k}{\sigma}\right) \tag{7.1}$$

Solve (7.1) by separation of variables, subject to the conditions

$$y(0,t) = 0, \quad y(L,t) = 0,$$
$$y(x,0) = f(x), \quad y_t(x,0) = 0,$$

leaving expansion coefficients in integral form. Summarize, in words, the effect(s) of the spring term by in (7.1).

8. (*Constant forcing function*) We saw in Section 19.1 that if the effects of gravity are included, then the governing PDE is

$$c^2 y_{xx} = y_{tt} + g. \tag{8.1}$$

That is, the PDE $L[y] = c^2 y_{xx} - y_{tt} - g$ is *non*homogeneous. Solve (8.1) subject to the conditions

$$y(0,t) = 0, \quad y(L,t) = 0,$$
$$y(x,0) = f(x), \quad y_t(x,0) = 0,$$

leaving expansion coefficients in integral form. HINT: The form $y(x,t) = X(x)T(t)$ gives

$$\frac{X''}{X} = \frac{1}{c^2}\frac{T''}{T} + \frac{g}{c^2 XT}. \tag{8.2}$$

Because of the last term in (8.2), which contains both x and t dependence, we are unable to successfully complete the separation process (i.e., we are unable to get all of the x dependence on one side of the equation and all of the t dependence on the other). Thus, we suggest seeking y in the form

$$y(x,t) = y_p(x) + X(x)T(t) \tag{8.3}$$

instead. Putting (8.3) into (8.1), obtain

$$c^2 y_p'' + c^2 X''T = XT'' + g. \tag{8.4}$$

Thus, we can remove the unwelcome g term by setting $c^2 y_p'' = g$. Then we can complete the separation of variable in (8.4) as usual. Mathematically, $y_p(x)$ is a **particular solution** of the nonhomogeneous equation (8.1) since it satisfies the full equation (7.1), and XT is a solution of the associated homogeneous equation $c^2 y_{xx} = y_{tt}$. But in physical terms you will find that it is simply the "static sag" of the string due to gravity, satisfying the problem

$$c^2 y_p''(x) = g, \quad y_p(0) = 0, \quad y_p(L) = 0. \tag{8.5}$$

9. (*Nonconstant forcing function*) In Exercise 8 we included a forcing term that was a constant. The suggested solution technique would have worked even if the forcing term were a nonconstant function of x. But in this exercise we allow for t dependence as well. Thus, consider the problem

$$\begin{gathered} c^2 y_{xx} = y_{tt} + F(x,t), \\ y(0,t) = 0, \quad y(L,t) = 0, \\ y(x,0) = f(x), \quad y_t(x,0) = 0. \end{gathered} \tag{9.1}$$

To solve, we can use essentially the same **eigenvector expansion method** that is used in Section 11.3.2 to solve the nonhomogeneous matrix problem $\mathbf{Ax} = \lambda\mathbf{x} + \mathbf{c}$, and again in Exercise 17 of Section 18.3 to solve the nonhomogeneous diffusion equation $\alpha^2 u_{xx} = u_t - F(x,t)$.

(a) Accordingly, solve (9.1) by seeking

$$y(x,t) = \sum_{n=1}^{\infty} b_n(t) \sin\frac{n\pi x}{L} \tag{9.2}$$

and expanding

$$F(x,t) = \sum_{n=1}^{\infty} F_n(t) \sin\frac{n\pi x}{L} \tag{9.3}$$

and

$$f(x) = \sum_{n=1}^{\infty} f_n \sin\frac{n\pi x}{L},$$

where the coefficients

$$F_n(t) = \frac{2}{L}\int_0^L F(x,t)\sin\frac{n\pi x}{L}\,dx \tag{9.4}$$

and

$$f_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx \tag{9.5}$$

are considered as known [i.e., computable from $F(x,t)$ and $f(x)$]. With $\omega_n = n\pi c/L$, show that

$$y(x,t) = \sum_{n=1}^{\infty}\left[f_n\cos\omega_n t + \frac{1}{\omega_n}\int_0^t F_n(\tau)\sin\omega_n(\tau - t)\,d\tau\right]\sin\frac{n\pi x}{L}, \tag{9.6}$$

(b) With the help of the Leibniz rule formally verify that (9.6) satisfies (9.1).

(c) Work out the solution (9.6) for the case where $F(x,t) = F_0 \sin\Omega t$ and $f(x) = 0$, assuming that the driving frequency does not equal any of the natural frequencies ω_n.

(d) Same as (c), but where Ω equals one of the natural frequencies, say ω_k.

10. (*Variable end conditions*) Thus far our boundary conditions have been constant in time. Here, we consider nonconstant conditions. Consider the problem

$$\begin{gathered} c^2 y_{xx} = y_{tt}, \\ y(0,t) = p(t), \quad y(L,t) = q(t), \\ y(x,0) = f(x), \quad y_t(x,0) = 0. \end{gathered} \tag{10.1}$$

Changing dependent variables from $y(x,t)$ to $z(x,t)$ according to

$$y(x,t) = z(x,t) + \left(1 - \frac{x}{L}\right)p(t) + \frac{x}{L}q(t), \tag{10.2}$$

show that the problem governing $z(x,t)$ is of the type treated in Exercise 9. NOTE: Notice how an "input" can be moved – from the boundary conditions to a forcing term in the PDE. In the present case the PDE on y was homogeneous and the boundary conditions were nonhomogeneous. Following the change of variables you should find that the PDE on z is nonhomogeneous and the boundary conditions are homogeneous.

11. The voltage $v(x,t)$ in an underground cable is governed by a PDE of the form

$$v_{xx} = Av_{tt} + Bv_t + Cv, \tag{11.1}$$

where A, B, C are constants. It is proposed that if v_1 is a solution of $v_{xx} = Av_{tt} + Bv_t$ and v_2 is a solution of $v_{xx} = Av_{tt} + Cv$, then $v = v_1 + v_2$ is, by superposition, a solution of (10.1). Give a critical evaluation of that proposal.

12. (*Longitudinal waves in a rod*) First, read Exercise 2 of Section 19.1. Consider a rod of length L, cross-sectional area A, Young's modulus E, and mass per unit length σ. At $x = 0$ the rod is attached to a rigid wall and at $x = L$ the rod is free. Prior to time $t = 0$ we pull on the free end with a force F_0, so the rod is in static equilibrium, with a uniform stress $s_0 = F_0/A$. At $t = 0$ we remove the force.

(a) Show that the problem governing the displacement is

$$\begin{aligned} c^2 u_{xx} &= u_{tt}, \\ u(0,t) = 0, &\quad u_x(L,t) = 0, \\ u(x,0) = \frac{s_0}{E}x, &\quad u_t(x,0) = 0. \end{aligned} \tag{12.1}$$

(b) Solve (12.1) for $u(x,t)$.
(c) Determine the stress at the wall, $s(0,t)$.

13. (*Uniqueness*) In this section and in the preceding exercises we have developed solution techniques for wave problems, most of which are of the form

$$\begin{aligned} c^2 y_{xx} &= y_{tt} + F(x,t), \\ y(0,t) = p(t), &\quad y(L,t) = q(t), \\ y(x,0) = f(x), &\quad y_t(x,0) = g(x). \end{aligned} \tag{13.1}$$

Show that the solution to (13.1) is unique. HINT: As usual in uniqueness proofs, let and $y_1(x,t)$ and $y_2(x,t)$ be two solutions and consider the difference $w(x,t) = y_1(x,t) - y_2(x,t)$. Show that w satisfies the homogeneous version of (13.1),

$$\begin{aligned} c^2 w_{xx} &= w_{tt}, \\ w(0,t) = 0, &\quad w(L,t) = 0, \\ w(x,0) = 0, &\quad w_t(x,0) = 0. \end{aligned} \tag{13.2}$$

Considering the integral

$$I(t) = \int_0^L (w_t^2 + c^2 w_x^2)\,dx, \tag{13.3}$$

show, with the help of Leibniz differentiation and (13.2), that $dI/dt = 0$. Thus, show that $I(t) = 0$ for all $t \geq 0$ and hence that $w(x,t) = 0$ for all $0 < x < L$ and $0 < t < \infty$. Since $w(x,t) = y_1(x,t) - y_2(x,t) = 0$, y_1 and y_2 are necessarily identical, so there exists at most one solution to (13.1).

14. (*Vibrating beam*) It is known from mechanics that the free vibration of a uniform beam is governed by the fourth-order PDE*

$$y_{xxxx} + \frac{\sigma}{EI} y_{tt} = 0, \tag{14.1}$$

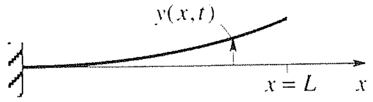
where $y(x,t)$ is the deflection, σ is the mass per unit length, E is Young's modulus of the material, and I is the moment of inertia of the cross section about its neutral axis.

(a) If the beam is "cantilevered" (see figure), then the boundary conditions are

$$\begin{aligned} &y(0,t) = 0, \quad y_x(0,t) = 0, \\ &y_{xx}(L,t) = 0, \quad \text{(no bending moment at free end)} \\ &y_{xxx}(L,t) = 0, \quad \text{(no shear force at free end)} \end{aligned} \tag{14.2}$$

and if the beam is initially deflected and at rest, then

$$y(x,0) = f(x), \quad y_t(x,0) = 0. \tag{14.3}$$



Seeking $y(x,t) = X(x)T(t)$, derive the solution form

$$y(x,t) = \sum_{n=1}^{\infty} A_n X_n(x) \cos \omega_n t, \tag{14.4}$$

where the mode shapes are given by

$$\begin{aligned} X_n(x) &= \sin \frac{z_n x}{L} - \sinh \frac{z_n x}{L} \\ &+ \left(\frac{\cos z_n + \cosh z_n}{\sin z_n - \sinh z_n} \right) \left(\cos \frac{z_n x}{L} - \cosh \frac{z_n x}{L} \right), \end{aligned} \tag{14.5}$$

where the frequencies are

[*] See for example, William T. Thomson, *Theory of Vibration with Applications*, 2nd ed. (Englewood Cliffs, NJ: Prentice Hall, 1981)

$$\omega_n = \sqrt{\frac{EI}{\sigma}}\left(\frac{z_n}{L}\right)^2, \tag{14.6}$$

and where the z_n's are the solutions of the transcendental equation $\cos z \cosh z + 1 = 0$. NOTE: The initial condition

$$y(x,0) = f(x) = \sum_{n=1}^{\infty} A_n X_n(x) \tag{14.7}$$

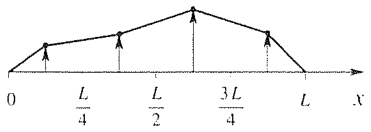
can then be used to evaluate the A_n's, but we do not ask you to go that far. Indeed, the eigenvalue problem on $X(x)$, which yields the eigenfunctions $X_n(x)$, is not of Sturm–Liouville type, so we do not find (in this text) the theoretical foundation needed to guide us through the evaluation of the A_n's in (14.7). Let it suffice to observe that (14.4) gives $y(x,t)$ as the superposition of infinitely many modes, having shape $X_n(x)$ and frequency ω_n.

(b) If, instead of being cantilevered the beam is pinned at both ends, then the boundary conditions are

$$\begin{aligned} y(0,t) = 0, \quad y(L,t) = 0, \\ y_{xx}(0,t) = 0, \quad y_{xx}(L,t) = 0, \end{aligned} \tag{14.8}$$

in place of (14.2). Find the solution form analogous to the one cited in part (a), corresponding to the boundary conditions (14.8).

15. (*Lumped-parameter model*) It is sometimes useful to model a *continuous* system, such as the string in our vibrating string problem, approximately, by a *discrete* system. To illustrate, let us divide the string into four equal parts and focus the mass $\sigma L/4$ of each segment at the center of that segment. (Such a system is called a **lumped-parameter** system.) Thus, we have four "beads" of mass $\sigma L/4$ connected by massless string under tension τ.



(a) Applying Newton's second law of motion to the first bead, show that

$$\frac{\sigma L}{4}\ddot{y}_1 \approx \tau\left(\frac{y_2 - y_1}{L/4} - \frac{y_1}{L/8}\right), \tag{15.1}$$

where dots denote time derivatives. Doing the same for the remaining three masses, show that the resulting ODE's can be expressed in matrix form as

$$\ddot{\mathbf{y}} + \mathbf{A}\mathbf{y} = \mathbf{0}, \tag{15.2}$$

where $\mathbf{y} = [y_1(t), \ldots, y_4(t)]^{\mathrm{T}}$ and

$$\mathbf{A} = \frac{16\tau}{\sigma L^2}\begin{bmatrix} 3 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 3 \end{bmatrix}. \tag{15.3}$$

(b) More generally, with N beads one obtains (15.2), where $\mathbf{y} = [y_1(t), \ldots, y_N(t)]^{\mathrm{T}}$ and

$$\mathbf{A} = \left(\frac{N}{L}\right)^2 \frac{\tau}{\sigma}\begin{bmatrix} 3 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & & -1 & 2 & -1 \\ 0 & \cdots & & 0 & -1 & 3 \end{bmatrix}. \tag{15.4}$$

(You need not derive this result.) Seeking a solution in the form

$$\begin{aligned} y_1(t) &= \eta_1 \sin(\omega t + \phi) \\ &\vdots \\ y_N(t) &= \eta_N \sin(\omega t + \phi) \end{aligned} \tag{15.5}$$

or $\mathbf{y}(t) = \boldsymbol{\eta}\sin(\omega t + \phi)$, show that

$$\mathbf{B}\boldsymbol{\eta} = \lambda\boldsymbol{\eta}, \tag{15.6}$$

where $\mathbf{B}$ is the matrix in (15.4), without the $(N/L)^2(\tau/\sigma)$ factor, and $\lambda = (\sigma/\tau)(L/N)^2\omega^2$.

(c) If, for a chosen value of N, one solves the eigenvalue problem (15.6), then one obtains approximations of the first N orthogonal mode shapes and eigenfrequencies of the continuous string: the eigenvectors $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N$ give the approximate mode shapes and the eigenvalues $\lambda_1, \ldots, \lambda_N$ give the approximate eigenfrequencies $\omega_1 = (N/L)\sqrt{\tau/\sigma}\sqrt{\lambda_1}, \ldots, \omega_N = (N/L)\sqrt{\tau/\sigma}\sqrt{\lambda_N}$. Here is the problem: Use computer software to solve (15.6) for the case where $N = 4$. Compare the computed eigenfrequencies with the exact values $\omega_1 = \pi c/L = (\pi/L)\sqrt{\tau/\sigma}, \ldots, \omega_N = N\pi c/L = (N\pi/L)\sqrt{\tau/\sigma}$, and compare the m computed mode shapes with the exact shapes $\sin(\pi x/L), \ldots, \sin(4\pi x/L)$.

(d) Same as (c), with $N = 10$.

19.3 Separation of Variables; Vibrating Membrane

In Section 19.1 we derive not only the equation $c^2 y_{xx} = y_{tt}$ governing the vibrating string but also the equation

$$\boxed{c^2(w_{xx} + w_{yy}) = w_{tt}} \tag{1a}$$

governing the vibrating membrane; $w(x, y, t)$ is the membrane deflection normal to the x, y plane, and $c^2 = \tau/\sigma$ where τ is the (uniform) tension per unit length and σ is the (uniform) mass per unit area. Let us consider the domain to be the rectangle $0 < x < a$ and $0 < y < b$ (Fig. 1), and let us solve (1) subject to the boundary conditions

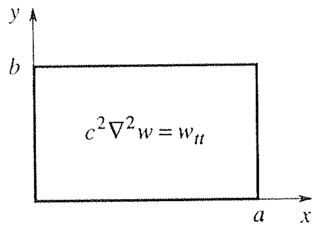


Figure 1. Rectangular membrane.

$$w(0, y, t) = w(a, y, t) = w(x, 0, t) = w(x, b, t) = 0 \tag{1b}$$

and the initial conditions

$$w(x, y, 0) = f(x, y), \quad w_t(x, y, 0) = 0, \tag{1c}$$

where $w(x, y, 0) = f(x, y)$ is the initial deflection and $w_t(x, y, 0)$ is the initial velocity. That is, the membrane is initially at rest.

The reason that we isolate this problem in a separate section is that it is our first problem with more than two independent variables – namely x, y, and t. Our approach here is intended to serve as a model for such cases.

To solve by separation of variables, seek

$$\boxed{w(x, y, t) = X(x)Y(y)T(t).} \tag{2}$$

Putting (2) into (1a) gives

$$c^2(X''YT + XY''T) = XYT''$$

or, dividing by c^2XYT,

$$\frac{X''}{X} + \frac{Y''}{Y} = \frac{1}{c^2}\frac{T''}{T}. \tag{3}$$

The left-hand side is a function of x and y, and the right-hand side is a function of t. Since x, y, t are independent variables, it follows in the usual way that each side must be a constant. Thus,

$$\frac{X''}{X} + \frac{Y''}{Y} = \frac{1}{c^2}\frac{T''}{T} = \text{constant} \equiv -\kappa^2, \tag{4}$$

with the minus sign included so that the T equation

$$T'' + \kappa^2 c^2 T = 0 \tag{5}$$

gives oscillatory solutions ($\cos \kappa ct$ and $\sin \kappa ct$), since we anticipate a vibratory motion. Next, we separate the x and y dependence in (4) by writing

$$\frac{X''}{X} = -\frac{Y''}{Y} - \kappa^2.$$

The left-hand side is a function of x alone and the right-hand side is a function of y alone so it follows, as usual, that

$$\frac{X''}{X} = -\frac{Y''}{Y} - \kappa^2 = \text{constant} \equiv -\alpha^2. \tag{6}$$

Hence,

$$X'' + \alpha^2 X = 0, \tag{7}$$

$$Y'' + (\kappa^2 - \alpha^2)Y = 0. \tag{8}$$

From (5), (7), and (8), we obtain

$$X = A\cos \alpha x + B \sin \alpha x, \tag{9a}$$

$$Y = D\cos\sqrt{\kappa^2 - \alpha^2}\, y + E\sin\sqrt{\kappa^2 - \alpha^2}\, y, \tag{9b}$$

$$T = F\cos \kappa ct + G\sin \kappa ct. \tag{9c}$$

The right-hand side of (9c) is the general solution of (5) only if $\kappa \neq 0$, for if $\kappa = 0$ then the sine term drops out. Consistent with the strategy that we have used until now, we should write, in place of (9c),

$$T = \begin{cases} F\cos \kappa ct + G\sin \kappa ct, & \kappa \neq 0 \\ H + It, & \kappa = 0. \end{cases} \tag{10}$$

Similarly for (9a) for the case $\alpha = 0$ and for (9b) for the case $\kappa = \alpha$. However, in this problem we anticipate that the I in (10) will be found to be zero when we apply the boundary and initial conditions because it gives a linear variation in t, whereas we expect an oscillatory motion. Similarly, we do not expect to need the linear terms in x and y (corresponding to the special cases $\alpha = 0$ and $\kappa = \alpha$) because $w = 0$ on the boundary. Thus, let us proceed with the solution forms (9a,b,c) rather than carry extra terms, terms that we know will drop out later.

Next, we put (9a,b,c) into (2) and apply the boundary and initial conditions. However, it is more efficient to observe from the boundary condition

$$w(0, y, t) = 0 = X(0)Y(y)T(t) \tag{11}$$

that $X(0) = 0$, and to observe from the other boundary conditions in (1b) that $X(a) = 0$, $Y(0) = 0$, and $Y(b) = 0$.* Applying the boundary condition $X(0) = 0$

*Alternatively, (11) is satisfied by $Y(y) = 0$ or $T(t) = 0$, but we cannot tolerate these choices because they give $w(x, y, t) = 0$, which will not satisfy the initial condition $w(x, y, 0) = f(x, y)$. That is, as usual, we make such choices so as to maintain as robust a solution as possible.

to (9a) gives $A = 0$, so $X(x) = B \sin \alpha x$. Then, $X(a) = 0 = B \sin \alpha a$. We cannot afford the choice $B = 0$ for then $X(x) = 0$ and $w(x, y, z) = 0$, so we require that $\sin \alpha a = 0$. Hence,

$$\alpha = \frac{m\pi}{a} \tag{12}$$

for any integer $m = 1, 2, \ldots$. Similarly, $Y(0) = 0$ gives $D = 0$ and $Y(b) = 0$ gives $\sqrt{\kappa^2 - \alpha^2} = n\pi/b$ for any integer $n = 1, 2, \ldots$, or because of (12),

$$\kappa = \pi\sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}. \tag{13}$$

Further, the initial condition

$$w_t(x, y, 0) = 0 = X(x)Y(y)T'(0) \tag{14}$$

gives $T'(0) = 0$, and application of this condition to (9c) gives $G = 0$.

Putting these results into (9a,b,c) and then putting the latter into (2) gives the product solution

$$w(x, y, t) = BEF \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \cos \pi c\sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}\, t \tag{15}$$

for any $m = 1, 2, \ldots, n = 1, 2, \ldots$, where $BEF \equiv H$, say, is arbitrary. Or, with the help of superposition,

$$\boxed{w(x, y, t) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} H_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \cos \omega_{mn} t,} \tag{16a}$$

where

$$\omega_{mn} = \pi c\sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}. \tag{16b}$$

Finally, the initial condition $w(x, y, 0) = f(x, y)$ requires that

$$\boxed{f(x, y) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} H_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}} \tag{17}$$

on the rectangular domain. The latter series is an example of a **double series**; more specifically, it is a **double Fourier series**.* Suppose that $f(x, y)$ can, indeed, be

*Recall that for "single series" we say that $\sum_{n=1}^{\infty} a_n$ converges to s if to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N_0(\epsilon)$ such that

$$\left| s - \sum_{n=1}^{\infty} a_n \right| < \epsilon$$

represented by such a series. Then we can compute the H_{mn} coefficients formally as follows.

Re-express (17) as

$$f(x,y) = \sum_{n=1}^{\infty} \left(\sum_{m=1}^{\infty} H_{mn} \sin \frac{m\pi x}{a} \right) \sin \frac{n\pi y}{b}$$
$$\equiv \sum_{n=1}^{\infty} R_n(x) \sin \frac{n\pi y}{b}. \tag{18}$$

For fixed x $(a < x < b)$ the latter is a half-range sine expansion of $f(x,y)$ on $0 < y < b$, so

$$R_n(x) = \frac{2}{b} \int_0^b f(x,y) \sin \frac{n\pi y}{b}\, dy. \tag{19}$$

In turn,

$$R_n(x) = \sum_{m=1}^{\infty} H_{mn} \sin \frac{m\pi x}{a} \tag{20}$$

is a half-range sine expansion of $R_n(x)$ on $0 < x < a$, so

$$H_{mn} = \frac{2}{a} \int_0^a R_n(x) \sin \frac{m\pi x}{a}\, dx. \tag{21}$$

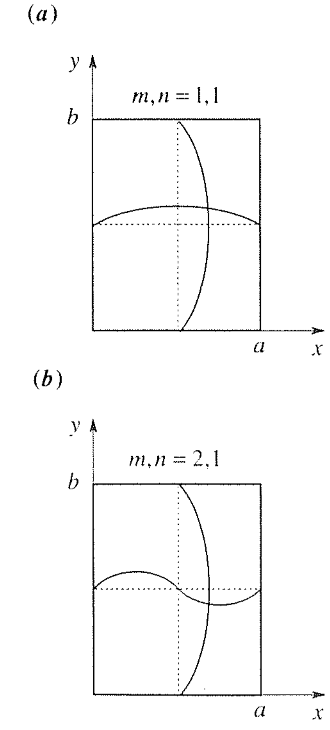


Figure 2. 1, 1 and 2, 1 mode shapes.

Putting (19) into (21) gives the formula

$$\boxed{H_{mn} = \frac{4}{ab} \int_0^b \int_0^a f(x,y) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}\, dx\, dy} \tag{22}$$

for the evaluation of the Fourier coefficients H_{mn} in (17). Thus, the solution of (1) is given by (16) with the coefficients calculated according to (22). With these results in hand let us comment on their derivation and physical interpretation.

COMMENT 1. Our choice of the minus sign in (4) was dictated by our understanding that the motion is, indeed, a vibration; the minus sign in (4) led to a plus sign

whenever $N > N_0$. Other definitions are used occasionally; this one is called **ordinary convergence** and we say that $\sum_{n=1}^{\infty} a_n$ converges to s in the sense of ordinary convergence. If to each $\epsilon > 0$ there does not exist such an then the series is said to diverge. Analogously, the **double series** $\sum_{n=1}^{\infty} \sum_{m=1}^{\infty} a_{mn}$ is said to converge to s, in the sense of *Pringsheim convergence*, if to each $\epsilon > 0$ (no matter how small) there correspond integers $M_0(\epsilon)$ and $N_0(\epsilon)$ such that

$$\left| s - \sum_{n=1}^{N} \sum_{m=1}^{M} a_{mn} \right| < \epsilon$$

whenever $M > M_0$ and $N > N_0$; otherwise the series is said to diverge. For further discussion, we refer the interested reader to P. W. Berg and J. L. McGregor, *Elementary Partial Differential Equations*, prelim. ed. (San Francisco: Holden–Day, 1964), Sec. 10.3.

in (5) and the oscillatory solutions $\cos\kappa ct$ and $\sin\kappa ct$. But what motivated us to choose the minus in front of the α^2 in (6)? Once again, we need to look ahead. Specifically, the eventual Fourier series expansion of $f(x,y)$ dictates the need for cosine and sine solutions of (7) and (8). The $+\alpha^2$ in (7), which results from the $-\alpha^2$ in (6), does give $\cos\alpha x$ and $\sin\alpha x$ solutions. Further, observe that we wrote the Y equation (8) in the form $Y''+(\kappa^2-\alpha^2)Y=0$ rather than in the equally correct form $Y''-(\alpha^2-\kappa^2)Y=0$ in order to force the cosine and sine solutions given in (9b) (see Exercise 1).

COMMENT 2. What do the individual modes look like? The shape of the m,n mode is given by $\sin(m\pi x/a)\sin(n\pi y/b)$, which is modulated periodically in time by the $\cos\omega_{mn}t$ factor. For $m,n=1,1$ and $2,1$, for instance, the mode shapes are as indicated schematically in Fig. 2. In the case of the 2, 1 mode $w=0$ all along the line $x=a/2$ because the $\sin(2\pi x/a)$ factor is zero on that line. We call that line, or any curve within the domain and along which $w=0$ for all time, a **nodal line**, and show it as solid in Fig. 2a. If we simply show the nodal lines and indicate the positive and negative deflections by $+$'s and $-$'s, then the first several mode patterns are as shown in Fig. 3. Of course, the $+$'s and $-$'s alternate in time due to the $\cos\omega_{mn}t$ factor.

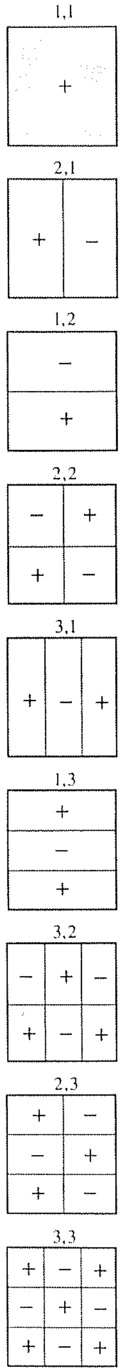


Figure 3. The first nine modes.

COMMENT 3. Concerning the temporal frequencies, observe that whereas for the vibrating string the frequencies $\omega_n=n\pi c/L$ are integer multiples of a fundamental frequency $\pi c/L$, the frequencies

$$\omega_{mn}=\pi c\sqrt{\frac{m^2}{a^2}+\frac{n^2}{b^2}}\ \frac{\text{rad}}{\text{sec}} \tag{23}$$

of the vibrating membrane are not. To illustrate, let $a=b$ and let us examine the musical notes corresponding to the various modes. For comparison with the analogous results for a violin string (Table 1, Section 19.2)), let us tune our "square drum" so that its fundamental frequency is 27.5 cycles/sec, corresponding to A_0, the lowest A on a piano. That is, adjust the tension τ so that

$$\omega_{1,1}=\left(\pi c\sqrt{\frac{1^2}{a^2}+\frac{1^2}{a^2}}\ \frac{\text{rad}}{\text{sec}}\right)\left(\frac{1\text{ cycle}}{2\pi\text{ rad}}\right)=\frac{c}{\sqrt{2}\,a}\ \frac{\text{cycles}}{\text{sec}}=27.5,$$

where $c=\sqrt{\tau/\sigma}$. Then the frequencies are

$$\begin{aligned}\omega_{mn}&=\left(\pi c\sqrt{\frac{m^2}{a^2}+\frac{n^2}{a^2}}\ \frac{\text{rad}}{\text{sec}}\right)\left(\frac{1\text{ cycle}}{2\pi\text{ rad}}\right)\\&=\frac{c}{2\,a}\sqrt{m^2+n^2}\ \frac{\text{cycles}}{\text{sec}}=27.5\sqrt{\frac{m^2+n^2}{2}}\ \frac{\text{cycles}}{\text{sec}}\end{aligned}$$

and these are listed in ascending order in Fig. 4, along with the corresponding musical note. From the tabulation in Fig. 4 we can see why square drums are not prized as musical instruments for, beginning with A_1, virtually every note is present. This profusion of notes is due to the dense values of $\sqrt{m^2+n^2}$. The result would be somewhat like playing the piano with our forearms rather than with our fingers. Circular drums are better and are discussed in the exercises.

mn	ω_{mn}	Note
11	$\underline{27.5}$	A_0
12,21	43.5	$\approx F_1$
22	$\underline{55.0}$	A_1
13,31	61.5	$\approx B_1$
23,32	70.1	$\approx C_2^{\#}$
14,41	80.2	$\approx E_2$
33	82.5	$\approx E_2$
24,42	87.0	$\approx F_2$
34,43	97.2	$\approx G_2$
15,51	99.2	$\approx G_2$
25,52	104.7	$\approx G_2^{\#}$
44	$\underline{110.0}$	A_2
35,53	113.4	$\approx A_2^{\#}$
16,61	118.3	$\approx A_2^{\#}$
26,62	123.0	$\approx B_2$
45,54	124.5	$\approx B_2$
36,63	130.4	$\approx C_3$
17,71,55	137.5	$\approx C_3^{\#}$
46,64	140.2	$\approx C_3^{\#}$

Figure 4. Square drum, octave overtones underlined.

COMMENT 4. Alternative to the product form (6) we can seek w in the product form

$$\boxed{w(x,y,t)=W(x,y)T(t).} \tag{24}$$

Putting (4) into (1a) and separating gives

$$c^2\left(W_{xx}+W_{yy}\right)T=WT'', \tag{25a}$$

$$\frac{W_{xx}+W_{yy}}{W}=\frac{1}{c^2}\frac{T''}{T}=-\kappa^2, \tag{25b}$$

and the separated equations

$$\boxed{W_{xx}+W_{yy}+\kappa^2W=0} \tag{26}$$

on W, and $T''+\kappa^2c^2T=0$ on T, as before. Then we can seek

$$W(x,y)=X(x)Y(y) \tag{27}$$

in (26). The resulting ODE's on X, Y, T are the same as before, but we mention this option for two reasons. First, some people prefer this method, and it is often used in textbooks. Second, (26) is a well known PDE, the **Helmholtz equation**, named after *Hermann von Helmholtz* (1821–1894) and studied by him in connection with acoustics.

EXAMPLE 1. Let the initial deflection be

$$w(x,y,0)=f(x,y)=0.2\sin\frac{\pi x}{a}\sin\frac{2\pi y}{b}-0.1\sin\frac{5\pi x}{a}\sin\frac{3\pi y}{b}. \tag{28}$$

We could put this f into (22) and integrate, to determine the H_{mn}'s, but it is much simpler to "match" terms in (17):

$$\begin{aligned}&0.2\sin\frac{\pi x}{a}\sin\frac{2\pi y}{b}-0.1\sin\frac{5\pi x}{a}\sin\frac{3\pi y}{b}\\&\quad=H_{11}\sin\frac{\pi x}{a}\sin\frac{\pi y}{b}+H_{12}\sin\frac{\pi x}{a}\sin\frac{2\pi y}{b}+H_{21}\sin\frac{2\pi x}{a}\sin\frac{\pi y}{b}+\cdots.\end{aligned} \tag{29}$$

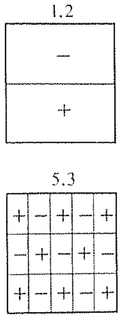


Figure 5. Modes in (30).

Thus, $H_{12}=0.2$, $H_{53}=-0.1$, and all other H_{mn}'s are zero. Then, ω_{12} and ω_{53} are obtained from (16b), so the solution is

$$\begin{aligned}w(x,y,t)&=0.2\sin\frac{\pi x}{a}\sin\frac{2\pi y}{b}\cos\left(\pi c\sqrt{\frac{1}{a^2}+\frac{4}{b^2}}\,t\right)\\&-0.1\sin\frac{5\pi x}{a}\sin\frac{3\pi y}{b}\cos\left(\pi c\sqrt{\frac{25}{a^2}+\frac{9}{b^2}}\,t\right).\end{aligned} \tag{30}$$

In this example only two modes are present, and these have nodal lines as indicated in Fig. 5. What, if any, are the nodal lines of the solution $w(x,y,t)$, lines along which $w=0$ for all t? Since the two cosine functions in (30) are linearly independent, $w=0$ for all t requires that both $\sin\frac{\pi x}{a}\sin\frac{2\pi y}{b}$ *and* $\sin\frac{5\pi x}{a}\sin\frac{3\pi y}{b}$ be zero. But the curves on which these products are zero are disjoint (Fig. 5), so the solution $w(x,y,t)$ has no nodal lines. In other cases $w(x,y,t)$ can, indeed, have nodal lines and these lines are not necessarily straight, as discussed in the exercises. ▮

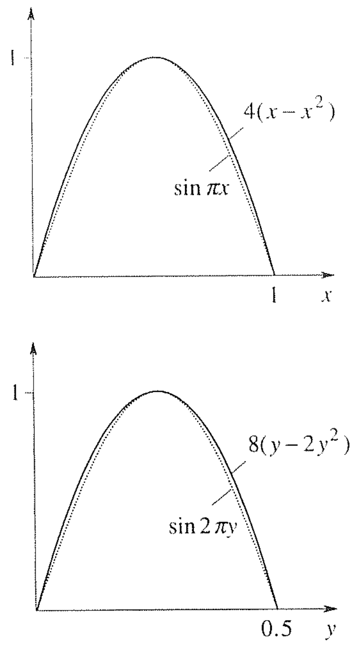


Figure 6. Similarity between the factors in (31) and the factors in the first mode.

EXAMPLE 2. Using the mks system of units (meters, kilograms, seconds, and newtons), let $a=1$ m, $b=0.5$ m, $\tau=8$ nt/m, $\sigma=0.02$ kg/m^2, and let the initial displacement be

$$w(x,y,0)=f(x,y)=0.02(x-x^2)(y-2y^2)\text{ meters.} \tag{31}$$

Then $c=\sqrt{\tau/\sigma}=20$ and, from (22),

$$\begin{aligned} H_{mn} &= 8\int_0^{0.5}\int_0^1 0.02(x-x^2)(y-2y^2)\sin m\pi x\sin 2n\pi y\,dx\,dy \\ &= 0.16\int_0^1 (x-x^2)\sin m\pi x\,dx\int_0^{0.5}(y-2y^2)\sin 2n\pi y\,dy \\ &= \frac{0.64}{m^3n^3\pi^6} \end{aligned} \tag{32}$$

if m and n are odd, and zero if m and/or n is even. Hence, the solution is

$$w(x,y,t)=\frac{0.64}{\pi^6}\sum_{n=1,3,\dots}^{\infty}\sum_{m=1,3,\dots}^{\infty}\frac{1}{m^3n^3}\sin m\pi x\sin 2n\pi y\cos\left(20\pi\sqrt{m^2+4n^2}\right)t. \tag{33}$$

COMMENT. Observe that the terms diminish rapidly with m and n due to the $1/(m^3n^3)$ factor, so that even the first term alone gives a reasonable approximation of the solution. That result is a consequence of the fact that the $(x-x^2)(y-2y^2)$ product in (31) is approximated well by a scalar multiple of the first mode $\sin\pi x\sin 2\pi y$ (Fig. 6). ▮

Closure. To solve the wave equation in three independent variables (x,y,t) by separation of variables we seek $w(x,y,t)$ in the form $X(x)Y(y)T(t)$. The separation process is successful, and we obtain ODE's on X,Y, and T, and two separation constants rather than one. The chief complication is that the solution is a double series rather a single series, with the coefficients given by a double integral rather than a single integral.

Computer software. Nodal lines are not always identified as readily as those in Fig. 3 but can be obtained using computer software. For instance, if $a=\pi$, $b=2\pi$, $c=1$, and

$$w(x,y,t)=(2\sin 3x\sin y-\sin x\sin 3y)\cos(\sqrt{10}\,t), \tag{34}$$

then the nodal lines, if any, are given implicitly by the relation

$$2\sin 3x\sin y-\sin x\sin 3y=0. \tag{35}$$

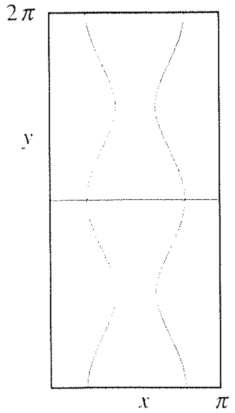


Figure 7. Nodal lines for (34).

Using *Maple*, enter

with (plots):

to access the plotting commands. Then use the **implicitplot** command as follows:

$$\text{implicitplot}\,(2 * \sin(3 * x) * \sin(y) - \sin(x) * \sin(3 * y) = 0,$$
$$x = 0..Pi, y = 2 * Pi, \text{grid} = [100, 200]);$$

The result is shown in Fig. 7. The "grid $= [100, 200]$" part is an option that we have used to set the plotting grid at 100 divisions of the x interval and 200 divisions of the y interval since w the default grid, $[25, 25]$, is too coarse in this case.

EXERCISES 19.3

1. Obtain the solution to the initial-value problem (1) for the given $f(x, y)$, a, and b, and give the relations that define the nodal lines of w, if any. Let $c = 1$ in each case.

(a) $f(x, y) = 8 \sin 2x \sin 2y, \quad a = \pi, b = 2\pi$
(b) $f(x, y) = -5 \sin x \sin 4y, \quad a = 2\pi, b = \pi$
(c) $f(x, y) = \sin 3x \sin y - \sin x \sin 3y, \quad a = b = \pi$
(d) $f(x, y) = \sin 3x \sin y - \sin x \sin 3y, \quad a = \pi, b = 2\pi$
(e) $f(x, y) = 1.05 \sin 3x \sin y - \sin x \sin 3y, \quad a = \pi, b = 2\pi$
(f) $f(x, y) = 10 \sin 3x \sin y - \sin x \sin 3y, \quad a = \pi, b = 2\pi$
(g) $f(x, y) = 8 \sin x \sin 7y + \sin 5x \sin 5y, \quad a = b = \pi$
(h) $f(x, y) = 8 \sin 2x \sin 7y + \sin 5x \sin 5y, \quad a = b = \pi$
(i) $f(x, y) = 6 \sin \pi x \sin 7\pi y + \sin 5\pi x \sin 5\pi y, \quad a = 1, b = 3$
(j) $f(x, y) = \sin \pi x \sin 4\pi y - \sin 3\pi x \sin 5\pi y, \quad a = 1, b = 2$

2. Obtain the solution to the initial-value problem (1) for the case where $f(x, y) = 20 \sin 3\pi x \sin 4\pi y - 8 \sin 5\pi x \sin 12\pi y$ and $a = b = c = 1$, and determine the period of the motion.

3. Evaluate the following claim and reasoning and indicate whether it is correct or incorrect. Claim: Whereas the solution to the vibrating string problem [given by (16) in Section 19.2] is periodic in time (for any choice of the R_n's and S_n's), the solution to the vibrating membrane problem (1) [given by (16a) in this section] is *not*, in general, a periodic function of t. Reasoning: It will suffice to consider two terms in (16a) so that $w(x, y)$ is of the form $A \cos \omega_{kl} t + B \cos \omega_{mn} t$ for some choice of the integers k, l, m, n and the constants A and B. (Of course, A and B contain x and y dependence, but we can consider x and y as fixed here.) If the period is T, then there must be integers M and N such that $\omega_{kl} T = (M)(2\pi)$ and $\omega_{mn} T = (N)(2\pi)$. Division gives $\omega_{kl}/\omega_{mn} = M/N$. That is, ω_{kl}/ω_{mn} must be a rational number. However, we see from (16b) that ω_{kl}/ω_{mn} is, except for certain choices of $k, l, m,$ and n, irrational.

4. (*Nonzero initial velocity*) We used $w_t(x, y, 0) = 0$ in (1c) merely for brevity.

(a) With

$$w(x, y, 0) = 0, \quad w_t(x, y, 0) = g(x, y)$$

in place of (1c), rederive the solution and obtain results analogous to (16) and (22).

(b) With

$$w(x, y, 0) = f(x, y), \quad w_t(x, y, 0) = g(x, y)$$

in place of (1c), re-derive the solution and obtain results analogous to (16) and (22).

5. (*Circular drum*) If the membrane is stretched over the circular disk $r < a$, then it is best to use the polar coordinates r, θ rather than the Cartesian coordinates x, y. Then the wave equation $c^2 \nabla^2 w = w_{tt}$ on $w(r, \theta, t)$ becomes

$$c^2 \left(w_{rr} + \frac{1}{r} w_r + \frac{1}{r^2} w_{\theta\theta} \right) = w_{tt}. \tag{5.1}$$

Let us consider only axisymmetric motions (i.e., we consider w to be independent of θ). Then the $w_{\theta\theta}$ term in (5.1) is zero, and (5.1) reduces to

$$c^2 \left(w_{rr} + \frac{1}{r} w_r \right) = w_{tt} \tag{5.2}$$

on $w(r, t)$.

(a) Seeking $w(r, t) = R(r) T(t)$, obtain

$$rR'' + R' + \kappa^2 r R = 0,$$
$$T'' + \kappa^2 c^2 T = 0$$

and hence the solution form

$$\begin{aligned}w(r,t) &= (A + B\ln r)(D + Et)\\ &+[FJ_0(\kappa r) + GY_0(\kappa r)](H\cos\kappa ct + I\sin\kappa ct).\end{aligned} \tag{5.3}$$

(b) If $w(a,t) = 0$ and $w(r,t)$ is to be bounded on $r < a$ (in particular at $r = 0$), show that we obtain

$$w(r,t) = \sum_{n=1}^{\infty} J_0\left(z_n \frac{r}{a}\right)(H_n\cos\omega_n t + I_n \sin\omega_n t), \tag{5.4}$$

where the z_n's ($n = 1, 2, \dots$) are the (known) zeros of J_0 [i.e., $J_0(z_n) = 0$] and $\omega_n = z_n c/a$.

(c) Let the initial conditions be

$$w(r,0) = f(r), \quad w_t(r,0) = g(r). \tag{5.5}$$

Imposing these conditions on (5.4), show that

$$\begin{aligned}H_n &= \frac{2}{a^2[J_1(z_n)]^2}\int_0^a f(r)J_0\left(z_n\frac{r}{a}\right) r\,dr,\\ I_n &= \frac{2}{\omega_n a^2[J_1(z_n)]^2}\int_0^a g(r)J_0\left(z_n\frac{r}{a}\right) r\,dr.\end{aligned} \tag{5.6a,b}$$

HINT: This problem is similar to Example 5 in Section 18.3.

(d) Show that if we seek $w(r,\theta,t) = R(r)\Theta(\theta)T(t)$ in the full equation (5.1), then we obtain the ODE's

$$\begin{aligned}r^2R'' + rR' + (\kappa^2r^2 - \alpha^2)R &= 0,\\ \Theta'' + \alpha^2\Theta &= 0,\\ T'' + \kappa^2c^2T &= 0,\end{aligned}$$

where κ and α are separation constants.

6. (*Two-dimensional diffusion*) In Chapter 18 we study one-dimensional diffusion phenomena such as the unsteady conduction of heat in a rod. Using the methods discussed in this section we can now return to Chapter 18 and solve two-dimensional problems as well. Specifically, consider the temperature field $u(x,y,t)$ in a rectangular plate ($0 < x < a$, $0 < y < b$), governed by the problem

$$\begin{aligned}\alpha^2(u_{xx} + u_{yy}) &= u_t,\\ u(0,y) = u(a,y) = u(x,0) &= u(x,b) = 0,\\ u(x,y,0) &= f(x,y).\end{aligned} \tag{6.1a,b,c}$$

(a) Derive the solution

$$u(x,y,t) = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} A_{mn}\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}e^{-\kappa_{mn}^2\alpha^2 t}, \tag{6.2}$$

where

$$\kappa_{mn}^2 = \pi^2\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right) \tag{6.3}$$

and

$$A_{mn} = \frac{4}{ab}\int_0^b\int_0^a f(x,y)\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}\,dx\,dy. \tag{6.4}$$

(b) Verify, formally, that (6.2)–(6.4) does satisfy (6.1).

(c) Evaluate the A_{mn}'s for the case where $f(x,y) = 100$.

19.4 Vibrating String; d'Alembert's Solution

19.4.1. d'Alembert's solution. For the wave equation

$$c^2y_{xx} = y_{tt} \tag{1}$$

there exists a striking solution form that is due to *Jean Le Rond d'Alembert* (1717–1783). D'Alembert's method is based on a change of independent variables, from x and t to ξ and η, say, according to the simple relations

$$\boxed{\begin{aligned}\xi &= x - ct,\\ \eta &= x + ct.\end{aligned}} \tag{2a,b}$$

To re-express (1) in terms of ξ and η, the needed "building blocks" are the $\partial/\partial x$ and $\partial/\partial t$ operators. According to the chain rule (Sections 13.4, 13.6.4),

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial x} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta} \tag{3}$$

and

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta}\frac{\partial \eta}{\partial t} = -c\frac{\partial}{\partial \xi} + c\frac{\partial}{\partial \eta}, \tag{4}$$

so (1) becomes

$$c^2\left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)\left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)y = \left(-c\frac{\partial}{\partial \xi} + c\frac{\partial}{\partial \eta}\right)\left(-c\frac{\partial}{\partial \xi} + c\frac{\partial}{\partial \eta}\right)y$$

or

$$\left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)(y_\xi + y_\eta) = \left(-\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)(-y_\xi + y_\eta)$$

or

$$y_{\xi\xi} + y_{\xi\eta} + y_{\eta\xi} + y_{\eta\eta} = y_{\xi\xi} - y_{\xi\eta} - y_{\eta\xi} + y_{\eta\eta}. \tag{5}$$

Assuming that y is well enough behaved so that $y_{\xi\eta} = y_{\eta\xi}$,* (5) simplifies to $4y_{\xi\eta} = 0$, or

$$y_{\xi\eta} = 0. \tag{6}$$

The point, then, is that when expressed in terms of ξ and η the wave equation (1) simplifies dramatically to the form (6), and we say that (6) is the **canonical**, or simplest, form of (1). It is simple because it can be solved directly by integration. First, integrating (6) with respect to η gives

$$y_\xi = \int 0\,\partial\eta = 0 + A(\xi) = A(\xi), \tag{7}$$

where the "constant of integration" A is allowed to be an arbitrary function of ξ since ξ was held fixed during the integration on η. Next, integration of (7) with respect to ξ gives $y = \int A(\xi)\,\partial\xi$ plus an arbitrary function of η, say $G(\eta)$. Since A is arbitrary we might as well simply write $F(\xi)$ in place of $\int A(\xi)\,\partial\xi$. Thus,

$$y = F(\xi) + G(\eta) \tag{8}$$

or, returning to the original variables,

$$\boxed{y(x,t) = F(x - ct) + G(x + ct),} \tag{9}$$

*See Theorem 13.3.1.

where F and G are arbitrary functions [although they do need to be twice differentiable if (9) is to satisfy (1)]. For example, each of the three functions

$$6e^{x-ct}, \qquad \sin(x-ct)+5\sin[3(x+ct)], \qquad 100-5(x-ct)^3-8e^{4(x+ct)^2} \tag{10}$$

is of the form (9) and it is readily verified that each one satisfies (1). To verify (9), in general, we can use the chain rule:

$$\begin{aligned} y_t &= F'(x-ct)\,\frac{\partial(x-ct)}{\partial t}+G'(x+ct)\,\frac{\partial(x+ct)}{\partial t} \\ &= -cF'(x-ct)+cG'(x+ct) \\ y_{tt} &= -cF''(x-ct)\,\frac{\partial(x-ct)}{\partial t}+cG''(x+ct)\,\frac{\partial(x+ct)}{\partial t} \\ &= c^2F''(x-ct)+c^2G''(x+ct). \end{aligned} \tag{11a}$$

Similarly, we obtain

$$y_{xx}=F''(x-ct)+G''(x+ct), \tag{11b}$$

so we see from (11) that (1) is satisfied. Remember that primes are standard notation for the derivative of a function of a single variable with respect to that variable; for functions of more than one variable we use the partial derivative notation. Thus, by $F'(x-ct)$, for example, we mean the ordinary derivative $\frac{dF(x-ct)}{d(x-ct)}$ of F with respect to its single argument $x-ct$. To illustrate, if

$$F(x-ct)=3\sin(x-ct)^2,$$

then

$$F'(x-ct)=6(x-ct)\cos(x-ct)^2.$$

We say that (9) is the **general solution** of the wave equation (1). There is no analog of (9) in Chapter 18; nowhere in Chapter 18 did we find the general solution of the diffusion equation. Even the infinite series solutions that we found by separation of variables, which contained an infinite number of arbitrary constants, are not general solutions; they are simply comprehensive enough so as to be capable of satisfying the initial condition.

It is interesting to compare the form of (9) with the form $y(x)=C_1y_1(x)+C_2y_2(x)$ of the general solution to a linear homogeneous second-order *ordinary* differential equation; in place of an arbitrary constant C_1 *times* $y_1(x)$ we have an arbitrary *function* of $x-ct$, and in place of an arbitrary constant C_2 *times* $y_2(x)$ we have an arbitrary *function* of $x+ct$.

To illustrate the use of (9), consider the infinite string problem

$$c^2y_{xx}=y_{tt}, \qquad (-\infty<x<\infty,\ 0<t<\infty) \tag{12a}$$

$$y(x,0)=f(x),\ \ y_t(x,0)=g(x). \qquad (-\infty<x<\infty) \tag{12b}$$

Since we already have in (9) the general solution of (12a), we can bypass (12a) and immediately impose the initial conditions (12b) in order to determine the functions F and G in (9). Doing so gives

$$y(x,0) = f(x) = F(x) + G(x), \tag{13a}$$
$$y_t(x,0) = g(x) = -cF'(x) + cG'(x) \tag{13b}$$

as two equations on F and G. Integrating (13b) from any fixed point, say 0, to x gives

$$\int_0^x g(\xi)\,d\xi = -cF(x) + cF(0) + cG(x) - cG(0), \tag{14}$$

and solving (13a) and (14) for $F(x)$ and $G(x)$ gives

$$F(x) = \frac{f(x)}{2} - \frac{1}{2c}\int_0^x g(\xi)\,d\xi + \frac{F(0)-G(0)}{2}, \tag{15a}$$
$$G(x) = \frac{f(x)}{2} + \frac{1}{2c}\int_0^x g(\xi)\,d\xi - \frac{F(0)-G(0)}{2}. \tag{15b}$$

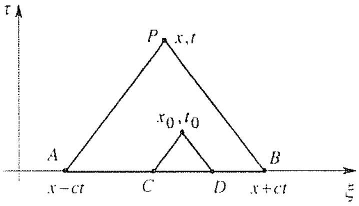


Figure 1. Interpretation of (17).

Since $F(x)$ is given by the right-hand side of (15a) $F(x-ct)$ is likewise given by the right-hand side of (15a), but with each of the two x's changed to $x-ct$. Similarly, $G(x+ct)$ is obtained by changing each of the x's in (15b) to $x+ct$. Doing so, we obtain

$$\begin{aligned} y(x,t) &= F(x-ct) + G(x+ct) \\ &= \frac{f(x-ct)}{2} - \frac{1}{2c}\int_0^{x-ct} g(\xi)\,d\xi + \frac{F(0)-G(0)}{2} \\ &\quad + \frac{f(x+ct)}{2} + \frac{1}{2c}\int_0^{x+ct} g(\xi)\,d\xi - \frac{F(0)-G(0)}{2}, \end{aligned} \tag{16}$$

or

$$\boxed{y(x,t) = \frac{f(x-ct)+f(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\xi)\,d\xi.} \tag{17}$$

Let us interpret (17) in the x,t plane. Since x,t is the specific point at which y is being evaluated, let us use dummy variables ξ, η for the axes (Fig. 1). In words, (17) tells us that y at P is the average of the f values at A and B plus $1/2c$ times the integral of g from A to B. Thus, the value of y at P depends on initial data only on the interval AB. Similarly, the value of y at x_0, t_0 depends on initial data only on the interval CD, and so on. Thus, we call the triangular region ABP the **domain of influence** of the interval AB; initial data on AB determine the solution within the triangle ABP.

This result for the wave equation is in contrast with the result that the solution $u(x,t)$ of the diffusion equation depends on the initial data $u(x,0) = f(x)$ all along the axis, as can be seen from (11) in Section 18.4 where the integration is from $\xi = -\infty$ to $\xi = +\infty$.

EXAMPLE 1. *The Case Where* $y_t(x,0) = 0$. To illustrate (17), suppose we release the string from rest in the configuration $y(x,0) = f(x)$, where f is the triangular pulse shown shown in Fig. 2. Then $g(x) = 0$ and (17) gives

$$y(x,t) = \frac{f(x-ct) + f(x+ct)}{2}. \tag{18}$$

At $t = 0$, (18) becomes $y(x,0) = f(x)/2 + f(x)/2$, so we may regard the initial pulse $f(x)$ as the sum of two "half-pulses," $f(x)/2$ and $f(x)/2$. For $t > 0$ the $f(x-ct)/2$ term in (18) amounts to one of the half-pulses translated to the right through a distance ct, and the $f(x+ct)/2$ term amounts to the other half-pulse translated to the left through a distance ct, as depicted in Fig. 3. Evidently, the speed of these right- and left-running waves is c. Thus, the $c = \sqrt{\tau/\sigma}$ that appears in (1) is the **wave speed**.

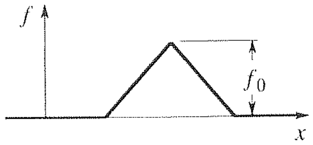


Figure 2. Initial shape of string: $y(x,0)f(x)$.

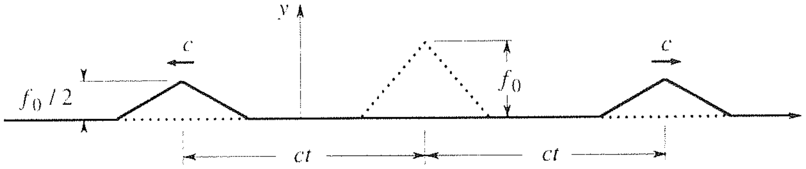


Figure 3. Right- and left-running waves.

Besides plotting y versus x as we have in Fig. 3, it is illuminating to display these results in the x, t plane as we have in Fig. 4. Naturally, plotting $y(x,t)$ above the x, t plane calls for three-dimensional graphics, but it is simpler to merely "lay the solution curves down" in the plane of the paper as we have in the figure. Since the $f(x-ct)/2$ term is constant along $x - ct =$ constant lines, the values of $f(x-ct)/2$ propagate along these lines without change. Similarly, the values of $f(x+ct)/2$ propagate along $x + ct =$ constant lines.

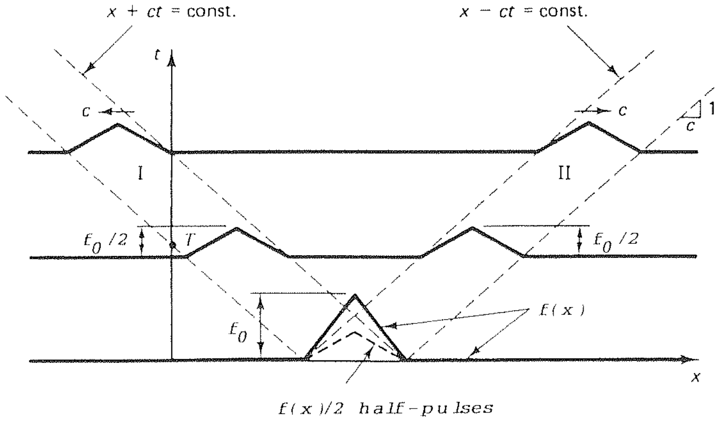


Figure 4. x, t plane display of (18).

These special families of curves, the lines $x - ct =$ constant and $x + ct =$ constant, are known as **characteristics**, and it is along the characteristics that information (i.e., values) propagates. In Fig. 4 we have drawn only four characteristics, those that define the "channels" I and II in which the half-pulses propagate, but there are actually an infinite number of characteristics, all of the curves $\xi = x - ct =$ constant and all of the curves $\xi = x + ct =$ constant as suggested more fully in Fig. 5.

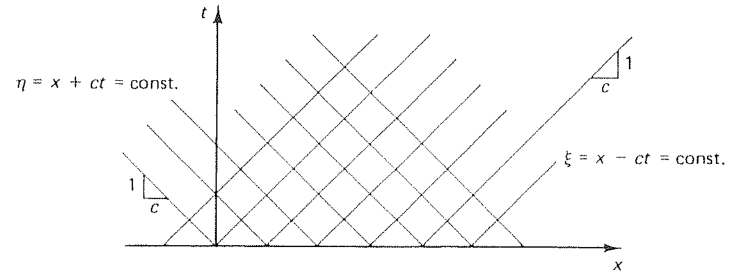


Figure 5. The two families of characteristics.

The upshot is that the initial f pulse breaks into two half-pulses that travel outward, without change in shape, in the channels I and II that are bounded by characteristics. Outside of these channels $y(x,t) = 0$. ▮

EXAMPLE 2. *The Case Where* $y(x,0) = 0$. (This example can be omitted if you have not read the optional Section 5.6 on the Dirac delta function.) In Example 1 we took $g(x) = 0$ and $f(x)$ to be a simple triangular pulse in order to gain understanding of the $[f(x-ct) + f(x+ct)]/2$ term in (17). Now, to study the integral term let the initial displacement be $y(x,0) = f(x) = 0$ and let the initial velocity be

$$y_t(x,0) = g(x) = \delta(x - x_0), \tag{19}$$

a delta function at some point x_0. These initial conditions are similar to those to which a piano wire is subjected, for the initial displacement is zero but a localized velocity is imparted at time $t = 0$ when the string is struck by a narrow hammer.

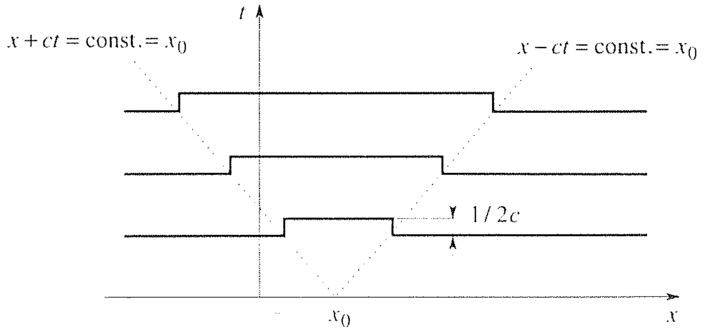


Figure 7. x, t plane display of (20).

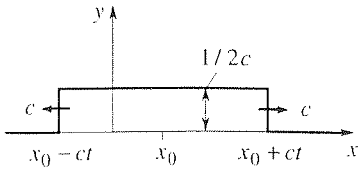


Figure 6. Response to hammer blow.

With $f(x) = 0$ and $g(x) = \delta(x - x_0)$, (17) gives

$$y(x,t) = 0 + \frac{1}{2c}\int_{x-ct}^{x+ct} \delta(\xi - x_0)\, d\xi. \tag{20}$$

From the definition of the delta function [see (13) in Section 5.6], the integral is 1 if $x - ct < x_0 < x + ct$ or, equivalently, if $x_0 - ct < x < x_0 + ct$. Hence, the solution (20) is the rectangular pulse shown in Fig. 6; in the x, t plane the solution is as shown in Fig. 7. ■

19.4.2. Use of images. (NOTE: The optional Section 18.5 on the method of images is not a prerequisite for this section.) Remember that the solution form (17) is for infinite strings, on $-\infty < x < \infty$. If we have a semi-infinite or finite string, then we have boundary conditions to deal with besides the two initial conditions. To illustrate, consider the following semi-infinite string problem:

$$c^2 y_{xx} = y_{tt}, \qquad (0 < x < \infty, \ 0 < t < \infty) \tag{21a}$$

$$y(0,t) = 0, \qquad (0 < t < \infty) \tag{21b}$$

$$y(x,0) = f(x), \ \ y_t(x,0) = 0, \qquad (0 < x < \infty) \tag{21c}$$

where f is the triangular pulse shown in Fig. 4. Observe that the solution shown in Fig. 4 does indeed satisfy (21) but only up until time T, at which time the left-running wave in channel I reaches the end point $x = 0$ and upsets the boundary condition (21b).

We can overcome this difficulty by an artifice known as the **method of images**. Namely, consider an *infinite* string ($-\infty < x < \infty$) with the initial conditions

$$y(x,0) = f_{\text{ext}}(x), \ \ y_t(x,0) = 0, \qquad (-\infty < x < \infty) \tag{22}$$

where $f_{\text{ext}}(x)$ is identical to $f(x)$ on $0 < x < \infty$ and is an antisymmetric extension of $f(x)$ for $-\infty < x < 0$ as shown in Fig. 8. The solution of the extended problem is

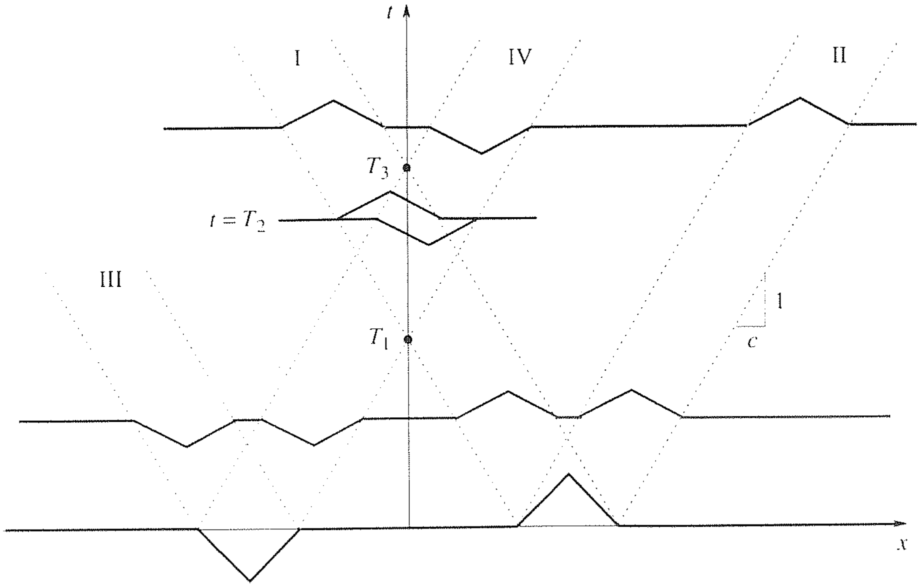


Figure 8. Image system.

$$y(x,0) = \frac{f_{\text{ext}}(x-ct) + f_{\text{ext}}(x+ct)}{2} \tag{23}$$

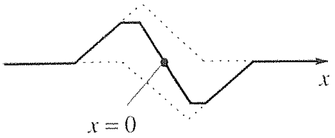


Figure 9. Cancellation at $x = 0$.

and the latter is comprised of two positive half-pulses in channels I and II plus the two negative half-pulses in channels III and IV. The point is that over the time interval $T_1 < t < T_3$ the waves in channels I and IV pass over each other, so that the negative wave in channel IV cancels the positive wave in channel I along the t axis. At $t = T_2$, for instance, the situation is as shown in Fig. 9, where the sum of the two waves is indicated by the solid line. That part of the diagram that is in the second quadrant $(-\infty < x < 0, 0 < t < \infty)$ is called the **image system** and is fictitious, serving only to automatically satisfy the boundary condition (21b) by virtue of the antisymmetry about $x = 0$. It can now be ignored, or even discarded, since the wave system in the first quadrant fully satisfies conditions (21a,b,c) and is the desired solution.

Since the image system is, after all, fictitious, to apprehend the physical event we look only at the first quadrant. We see that the initial pulse breaks into two half-pulses. One travels rightward indefinitely, while the left-running wave is both *reflected* (into channel *IV*) and *inverted* when it encounters the left end of the string, which is tied at $x = 0$.

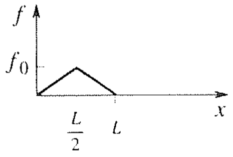


Figure 10. $f(x)$ in (24).

Having studied the infinite string and semi-infinite string, in that order, we can finally return to the finite string problem

$$c^2 y_{xx} = y_{tt}, \qquad (0 < x < L,\ 0 < t < \infty) \tag{24a}$$

$$y(0,t) = 0,\ \ y(L,t) = 0, \qquad (0 < t < \infty) \tag{24b}$$

$$y(x,0) = f(x),\ \ y_t(x,0) = 0 \qquad (0 < x < L) \tag{24c}$$

that is solved by separation of variables in Section 19.2, and solve it by d'Alembert's method – with help from the method of images. Let $f(x)$ be our usual generic triangular pulse (Fig. 10). Using the method of images, we consider instead the *infinite* string problem with initial conditions

$$y(x,0) = f_{\text{ext}}(x),\ \ y_t(x,0) = 0, \qquad (-\infty < x < \infty) \tag{25}$$

where $f_{\text{ext}}(x)$ is identical to $f(x)$ on $0 < x < L$ and is defined on $x < 0$ and on $x > L$ so as to be antisymmetric about *both* $x = 0$ and $x = L$. Thus, $f_{\text{ext}}(x)$ is the $2L$-periodic function shown in Fig. 11. Since the extended initial conditions (25) are antisymmetric about $x = 0$ and $x = L$, the solution $y(x,t)$ will be too. Hence, y will be zero all along the $x = 0$ and $x = L$ lines in the x,t plane. That is, the image system is designed so as to satisfy the boundary conditions (24b).

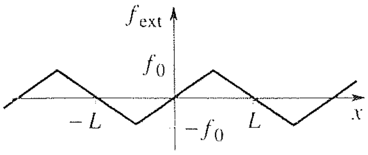


Figure 11. $f_{\text{ext}}(x)$, antisymmetric about $x = 0$ and $x = L$.

With $f_{\text{ext}}(x)$ as defined in Fig. 11, the d'Alembert solution (17) gives the closed form solution

$$y(x,t) = \frac{f_{\text{ext}}(x-ct) + f_{\text{ext}}(x+ct)}{2}. \tag{26}$$

To relate (26) to the separation of variables solution obtained in Section 19.2, expand the periodic function $f_{\text{ext}}(x)$ in a Fourier series as

$$f_{\text{ext}}(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}, \tag{27}$$

where

$$b_n = \frac{2}{L}\int_0^L f_{\text{ext}}(x)\sin\frac{n\pi x}{L}\,dx = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx$$

since $f_{\text{ext}}(x) = f(x)$ on $0 < x < L$. Then (26) becomes

$$\begin{aligned} y(x,t) &= \frac{1}{2}\sum_{n=1}^{\infty} b_n\left[\sin\frac{n\pi}{L}(x-ct) + \sin\frac{n\pi}{L}(x+ct)\right] \\ &= \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}\cos\frac{n\pi ct}{L}, \end{aligned} \tag{28}$$

which is identical to the result that we obtained by separation of variables.

19.4.3. Solution by integral transforms. (Optional) Following d'Alembert, we derived the solution (17) of the infinite string problem (12) by using the general solution $y(x,t) = F(x-ct) + G(x+ct)$. Alternatively, we could solve (12) by Laplace transforming with respect to the t variable or by Fourier transforming with respect to the x variable.

Let us try the Laplace transform. Transforming (12a) gives

$$\begin{aligned} c^2\overline{y}_{xx} &= s^2\overline{y} - sy(x,0) - y_t(x,0) \\ &= s^2\overline{y} - sf(x) - g(x) \end{aligned}$$

or

$$\overline{y}_{xx} - \frac{s^2}{c^2}\overline{y} = -\frac{1}{c^2}[sf(x) + g(x)]. \tag{29}$$

The homogeneous solution of (29) is simple but obtaining a particular solution (e.g., by the method of variation of parameters) is messy, so let us see whether the Fourier transform is more convenient. Fourier transforming (12a) gives

$$c^2(i\omega)^2\hat{y} = \hat{y}_{tt}$$

or

$$\hat{y}_{tt} + \omega^2c^2\hat{y} = 0, \tag{30}$$

which is simpler than (29) because it is homogeneous, so let us continue. From (30),

$$\hat{y} = A\cos\omega ct + B\sin\omega ct. \tag{31}$$

To solve for A and B we impose Fourier transformed versions of (12b):

$$\hat{y}\Big|_{t=0} = \hat{f}(\omega) = A, \tag{32a}$$

$$\hat{y}_t\Big|_{t=0} = \hat{g}(\omega) = \omega cB, \tag{32b}$$

so $A = \hat{f}(\omega)$ and $B = \hat{g}(\omega)/\omega c$ and

$$\hat{y} = \hat{f}(\omega)\cos\omega ct + \hat{g}(\omega)\frac{\sin\omega ct}{\omega c}. \tag{33}$$

To invert the two terms on the right, use entries 15, 9, and 21 of Appendix D and obtain

$$F^{-1}\left\{\hat{f}(\omega)\cos\omega ct\right\} = \frac{f(x-ct)+f(x+ct)}{2}, \tag{34a}$$

$$\begin{aligned} F^{-1}\left\{\hat{g}(\omega)\frac{\sin\omega ct}{\omega c}\right\} &= g(x) * \frac{1}{2c}[H(x+ct) - H(x-ct)] \\ &= \frac{1}{2c}\int_{-\infty}^{\infty} g(x-\xi)[H(\xi+ct) - H(\xi-ct)]\,d\xi \\ &= \frac{1}{2c}\int_{-ct}^{ct} g(x-\xi)\,d\xi \qquad (x-\xi=\mu) \\ &= \frac{1}{2c}\int_{x-ct}^{x+ct} g(\mu)\,d\mu, \end{aligned} \tag{34b}$$

where the third equality in (34b) follows from the fact that $H(x+ct) - H(x-ct)$ is zero for $x < -ct$, unity for $-ct < x < ct$, and zero for $x > ct$. Thus,

$$y(x,t) = \frac{f(x-ct)+f(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} g(\mu)\,d\mu,$$

which result is the same as (17).

Closure. Following d'Alembert, we change the independent variables from x, t to ξ, η according to the relations $\xi = x - ct$ and $\eta = x + ct$. That change reduces the wave equation $c^2 y_{xx} = y_{tt}$ to its canonical (i.e., simplest) form $y_{\xi\eta} = 0$, which can be solved by integration to give the general solution of the wave equation

$$y(x,t) = F(x-ct) + G(x+ct),$$

where F and G are arbitrary twice differentiable functions (Exercise 15). To illustrate the use of this general solution we solve an infinite string problem with prescribed initial displacement $f(x)$ and velocity $g(x)$ and obtain the solution (17).

The ξ, η variables are the most "natural" independent variables and the $\xi =$ constant and $\eta =$ constant lines constitute two families of lines (Fig. 5) called characteristics. Rather than kinks and discontinuities in the initial conditions smoothing out, as occurs in the process of diffusion, they propagate into the solution domain along characteristics. And whereas information spreads to the left and right at an infinite speed for the diffusion equation, it spreads with a finite wave speed c for the wave equation.

Remember that, while extremely important in their own right, the diffusion ($\alpha^2 u_{xx} = u_t$) and wave ($c^2 y_{xx} = y_{tt}$) equations also serve to represent the parabolic and hyperbolic types of PDE's, respectively, and the main features of these two equations are shared by other PDE's of their type. Thus, every hyperbolic PDE has its own families of characteristics, which may be curved rather than straight and along which information propagates.

In Section 19.4.2 we use the infinite string solution (17) together with the method of images to solve semi-infinite and finite string problems. If you studied Section 18.5 you will understand that the method of images works, here, because the $L = c^2\partial^2/\partial x^2 - \partial^2/\partial t^2$ operator is linear and even, but it is not necessary to understand *why* the method works in these cases inasmuch as we can *see* that it does lead to satisfaction of the homogeneous boundary conditions.

Finally, in Section 19.4.3 we rederive the solution (17), this time using a Fourier transform on x.

EXERCISES 19.4

1. Verify that (17) does satisfy (12a) and (12b).

2. Show that, like the wave equation, the given PDE is hyperbolic and find its general solution by introducing the suggested change of variables.

(a) $u_{xx} + 4u_{xy} + 3u_{yy} = 0$; $\xi = x - y$, $\eta = 3x - y$
(b) $u_{xx} - 4u_{xy} - 5u_{yy} = 0$; $\xi = x - y$, $\eta = 5x + y$
(c) $u_{xx} + 6u_{xy} + 8u_{yy} = 0$; $\xi = 4x - y$, $\eta = 2x - y$
(d) $u_{xx} + 4u_{xy} - 5u_{yy} = 0$; $\xi = x + y$, $\eta = 5x - y$
(e) $u_{xx} + 2u_{xy} - 3u_{yy} = 0$; $\xi = 3x - y$, $\eta = x + y$

3. Show that, like the wave equation, the given PDE is hyperbolic and find its general solution by introducing a suitable change of variables of the form $\xi = ax + by, \eta = cx + dy$.

(a) $u_{xx} + 8u_{xy} + 12u_{yy} = 0$
(b) $u_{xx} - 2u_{xy} - 3u_{yy} = 0$
(c) $u_{xx} - 10u_{xy} + 9u_{yy} = 0$
(d) $u_{xx} + 2u_{xy} - 8u_{yy} = 0$

4. Find the general solution of the first-order PDE $u_t + cu_x = 0$, where c is a constant, by introducing the change of variables $\xi = x - ct$, $\eta = t$, and then use that general solution to solve the problem.

$$\begin{aligned} &u_t + cu_x = 0, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \\ &u(x,0) = f(x). \end{aligned}$$

5. In Figs. 3 and 4 we show $y(x,t)$ at times t which were large enough so that the two half-pulses had completely separated (i.e., they do not overlap). In this exercise we examine the case in which they do overlap. Letting

$$f(t) = \begin{cases} 0, & x < 1 \\ 2x - 2, & 1 < x < 1.5 \\ 4 - 2x, & 1.5 < x < 2 \\ 0, & x > 2, \end{cases}$$

$g(x) = 0$, and $c = 20$, give labeled graphs of $y(x,t)$ at $t = 0.005$ and at $t = 0.02$.

6. With $c = 100$, sketch the solution to (12) at $t = 0.02$ and $t = 0.04$ using an x, t plane display similar to those in Figs. 4 and 7, and labeling all key values. As usual, H is the Heaviside function.

(a) $f(x) = H(x+1) - H(x-1)$, $g(x) = 0$
(b) $f(x) = H(x)$, $g(x) = 0$
(c) $f(x) = 0$, $g(x) = H(x+1) - H(x-1)$
(d) $f(x) = 0$, $g(x) = H(x)$

7. (a) Obtain the general solution for the nonhomogeneous wave equation $c^2 y_{xx} = y_{tt} - K$, where K is a constant.
(b) Use the general solution obtained in part (a) to solve the problem

$$\begin{aligned} &c^2 y_{xx} = y_{tt} - K, \qquad (-\infty < x < \infty,\ 0 < t < \infty) \\ &y(x,0) = p(x),\ y_t(x,0) = q(x). \end{aligned}$$

8. Use the general solution (9) to solve the problem

$$\begin{aligned} c^2 y_{xx} &= y_{tt}, \qquad (0 < x < \infty,\ 0 < t < \infty) \\ y(x,0) &= y_t(x,0) = 0. \qquad (0 < x < \infty) \\ y(0,t) &= h(t), \qquad (0 < t < \infty) \end{aligned}$$

where $h(t)$ is prescribed. [We can imagine taking the left end of the string between two fingers and, beginning at $t = 0$, "jiggling" it according to $y(0,t) = h(t)$.]

9. Same as Exercise 8, but solve by using the Laplace transform rather than (9). HINT: You may assume the truth of the statement

$$\lim_{x\to\infty} \overline{y}(x,s) = \lim_{x\to\infty} \int_0^\infty y(x,t)e^{-st}\,dt$$
$$= \int_0^\infty \lim_{x\to\infty} y(x,t)e^{-st}\,dt = \int_0^\infty 0\,dt = 0.$$

10. If spherical symmetry is present so that u depends only on ρ and t (where ρ, ϕ, θ are the spherical polar coordinates), then the wave equation $c^2\nabla^2 u = u_{tt}$ becomes

$$\frac{c^2}{\rho^2}\frac{\partial}{\partial \rho}\left(\rho^2 \frac{\partial u}{\partial \rho}\right) = \frac{\partial^2 u}{\partial t^2}. \tag{10.1}$$

Show that (10.1) may be re-expressed as

$$c^2 \frac{\partial^2}{\partial \rho^2}(\rho u) = \frac{\partial^2}{\partial t^2}(\rho u) \tag{10.2}$$

and thus derive the general solution

$$u(\rho,t) = \frac{1}{\rho}[F(\rho - ct) + G(\rho + ct)] \tag{10.3}$$

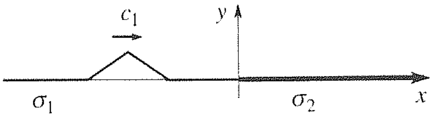
of (10.1), where F and G are arbitrary twice-differentiable functions. NOTE: Observe from the $1/\rho$ factor in (10.3) that in the case of spherical waves the wave amplitude tends to zero as $\rho \to \infty$.

11. Show whether traveling wave solutions of the form $F(x - at)$ are possible for the diffusion problem

$$\alpha^2 u_{xx} = u_t, \qquad (-\infty < x < \infty,\ 0 < t < \infty)$$
$$u(x,0) = f(x), \qquad (-\infty < x < \infty)$$

where $u(x,t)$ is bounded, for all t, as $x \to \pm\infty$.

12. (*Infinite string with density discontinuity*) Consider a string stretched over $-\infty < x < \infty$ under a tension τ, where the density $\sigma(x)$ has a step discontinuity at $x = 0$; i.e., $\sigma = \sigma_1$ for $x < 0$ and $\sigma = \sigma_2$ for $x > 0$. Suppose a rightward-running wave $y(x,t) = F(x - c_1 t)$, such as a triangular pulse, is initiated in the $x < 0$ part of the string, where $c_1 = \sqrt{\tau/\sigma_1}$ is the wave speed for $x < 0$ and $c_2 = \sqrt{\tau/\sigma_2}$ is the wave speed for $x > 0$. What happens when this wave encounters the density discontinuity (see the accompanying figure)? That is, determine the solution $y(x,t)$. HINT: Use the general solution $y(x,t) = G(x - c_1 t) + H(x + c_1 t)$ for $x < 0$, and $y(x,t) = I(x - c_2 t) + J(x + c_2 t)$ for $x > 0$. The solution will be nonzero only within certain channels, in which you need to determine the functions G, H, I, and J. You will need to match the two solutions (i.e., for $x < 0$ and for $x > 0$) along the line $x = 0$ in the x,t plane.



13. Use (17) to show that if $y(x,0) = f(x)$ and $y_t(x,0) = g(x)$ are even functions of x, then $y(x,t)$ remains an even function for all $t > 0$, and that if $f(x)$ and $g(x)$ are odd functions of x, then $y(x,t)$ remains an odd function for all $t > 0$.

14. We solve (21) by the method of images, where $f(x)$ is the triangular pulse shown in Fig. 4. Repeat that solution, with (21b) changed to $y_x(0,t) = 0$, and obtain an x,t plane diagram analogous to that in Fig. 8. Is there a reflection and an inversion as in Fig. 8? Explain.

15. (*Breakdowns at kinks*) We have emphasized that kinks and discontinuities in the initial conditions $y(x,0) = f(x)$ and $y_t(x,0) = g(x)$ propagate into the solution domain as in Fig. 4, for instance, where the kinks in f propagate along both right- and left-running characteristics. However, it must be confessed that at each point along those characteristics the PDE $c^2 y_{xx} = y_{tt}$ is not satisfied because both y_{xx} and y_{tt} fail to exist. (That y_{xx} does not exist there should be evident from Fig. 4. Do you see why y_{tt} fails to exist there as well?) Explain why results such as those displayed in Fig. 4 are acceptable nonetheless.

16. (*Finite-difference method*) In Section 18.6 we discretize the problem by means of the computational grid shown there in Fig. 1, and use difference quotient approximations of u_{xx} and u_t to obtain the computational formula $U_{j,k+1} = rU_{j-1,k} + (1 - 2r)U_{j,k} + rU_{j+1,k}$, where $U_{j,k}$ denotes the exact solution of the difference equation.

(a) Proceeding in the same manner, derive the scheme

$$Y_{i,j+1} = r^2 Y_{i-1,j} + 2(1 - r^2)Y_{i,j} + r^2 Y_{i+1,j} - Y_{i,j-1} \tag{16.1}$$

for the wave equation problem

$$\begin{gathered} c^2 y_{xx} = y_{tt}, \\ y(0,t) = p(t),\ \ y(L,t) = q(t), \\ y(x,0) = f(x),\ \ y_t(x,0) = g(x) \end{gathered} \tag{16.2,3,4}$$

on $0 < x < L$, $0 < t < \infty$, where $r = c\Delta t/\Delta x$.

(b) Letting $c = 10$, $L = 1$, $\Delta x = 0.25$, $\Delta t = 0.02$, $p(t) = q(t) = f(x) = 0$, and $g(x) = \sin \pi x$, use (16.1) to evaluate the $Y_{i,j}$'s along the first two time lines, i.e., at $(i, j) = (1,1), (2,1), (3,1), (1,2), (2,2), (3,2)$. Compare your results with the exact solution. HINT: For points on the first time (16.1) gives

$$Y_{1,1} = r^2 Y_{0,0} + 2(1 - r^2)Y_{1,0} + r^2 Y_{2,0} - Y_{1,-1},$$
$$Y_{2,1} = r^2 Y_{1,0} + 2(1 - r^2)Y_{2,0} + r^2 Y_{3,0} - Y_{2,-1},$$
$$Y_{3,1} = r^2 Y_{2,0} + 2(1 - r^2)Y_{3,0} + r^2 Y_{4,0} - Y_{3,-1},$$

but how are we to deal with the $Y_{1,-1}, Y_{2,-1}$ and $Y_{3,-1}$ terms when the points $(i, j) = (1, -1), (2, -1)$, and $(3, -1)$ do not lie on the computational grid? Use a finite-difference version of the initial condition $y_t(x, 0) = g(x)$.

(c) Same as (b) but with $p(t) = 20t$.

(d) Same as (b) but with $q(t) = 1000(1 - \cos t)$.

(e) Same as (b) but with $f(x) = 10 \sin \pi x$ and $g(x) = 0$.

(f) It can be shown that for the stability and convergence of the scheme (16.1) we need $r \leq 1$. You need not derive this result; we merely ask you to interpret it graphically in terms of the concept of the domain of influence.

Chapter 19 Review

The wave and diffusion equations are similar in several ways:

1. For both the one-dimensional wave and diffusion equations the solution domain is, typically, a semi-infinite strip ($0 < x < L$, $0 < t < \infty$), the quarter plane $0 < x < \infty$, $0 < t < \infty$, or the half plane $-\infty < x < \infty$, $0 < t < \infty$.

2. For both equations we can use the method of separation of variables, a Laplace transform on the t variable, or a Fourier transform on the x variable if the x domain is $-\infty < x < \infty$.

3. When solving by separation of variables, the boundary conditions are to be applied before the initial conditions.

But the wave and diffusion equations also differ in some ways:

1. A single initial condition $u(x, 0) = f(x)$ is appropriate for the diffusion equation $\alpha^2 u_{xx} = u_t$, whereas the two initial conditions $y(x, 0) = f(x)$ and $y_t(x, 0) = g(x)$ are appropriate for the wave equation $c^2 y_{xx} = y_{tt}$.

2. Only for the wave equation do we find a *general solution*, namely,

$$y(x, t) = F(x - ct) + G(x + ct), \tag{1}$$

where F and G are arbitrary (twice differentiable) functions. The graphs of $F(x - ct)$ and $G(x + ct)$, plotted versus x, translate rightward and leftward, respectively, with speed c. Of course, F and G need not be single terms. For example, in the solution

$$y(x, t) = \frac{1}{2} \sum_{n=1}^{\infty} a_n \sin \frac{n\pi}{L}(x - ct) + \frac{1}{2} \sum_{n=1}^{\infty} a_n \sin \frac{n\pi}{L}(x + ct) \tag{2}$$

of the problem

$$c^2 y_{xx} = y_{tt}, \qquad (-\infty < x < \infty,\ \ 0 < t < \infty) \tag{3a}$$

$$y(x,0) = f(x),\ \ y_t(x) = 0, \qquad (-\infty < x < \infty) \tag{3b}$$

where

$$a_n = \frac{2}{L}\int_0^L f(x)\,\sin\frac{n\pi x}{L}\,dx,$$

each of the functions

$$F(x-ct) = \frac{1}{2}\sum_{n=1}^{\infty} a_n \sin\frac{n\pi}{L}(x-ct), \tag{4a}$$

$$G(x+ct) = \frac{1}{2}\sum_{n=1}^{\infty} a_n \sin\frac{n\pi}{L}(x+ct) \tag{4b}$$

is a superposition of individual *traveling waves*. In this case the trigonometric identities $\sin(A \pm B) = \sin A \cos B \pm \sin B \cos A$ reduce (2) to the form

$$y(x,t) = \sum_{n=1}^{\infty} a_n \sin\frac{n\pi x}{L}\,\cos\frac{n\pi ct}{L}, \tag{5}$$

which is a superposition of *standing waves*. Thus, the wave equation inevitably gives traveling waves, by virtue of (1), and in some cases these traveling waves sum to standing waves. The variables $\xi = x - ct$ and $\eta = x + ct$ give two families of *characteristics* in the x,t plane, the $\xi =$ constant and $\eta =$ constant lines, along which information propagates as can be seen from (1) because $F(x-ct)$ is constant along $x - ct =$ constant lines and $G(x+ct)$ is constant along $x + ct =$ constant lines.

3. Whereas diffusion is a smoothing process, we see in Chapter 19 that kinks and discontinuities in the initial conditions propagate into the x,t solution domain indefinitely, along characteristics.

The *two*-dimensional wave equation

$$c^2(w_{xx} + w_{yy}) = w_{tt}, \tag{6}$$

such as governs the displacement $w(x,y,t)$ of a vibrating drumhead, can be solved by separation of variables by seeking

$$w(x,y,t) = W(x,y)T(t) \tag{7}$$

and obtaining

$$W_{xx} + W_{yy} + \kappa^2 W = 0, \tag{8a}$$

$$T'' + \kappa^2 c^2 T = 0, \tag{8b}$$

where (8a) is the two-dimensional *Helmholtz equation*. In turn, (8a) can be separated by seeking $W(x, y) = X(x)Y(y)$, or we could seek $w(x, y, t) = X(x)Y(y)T(t)$ right from the start. The result is a *double series*, and each initial condition $[w(x, y, 0) = f(x, y),\ w_t(x, y, 0) = g(x, y)]$ leads to a double Fourier series expansion. The same method can be applied to the two-dimensional diffusion equation $\alpha^2(u_{xx} + u_{yy}) = u_t$, the chief difference being that for the wave equation the T equation (8b) gives the oscillatory solutions $\cos \kappa ct$ and $\sin \kappa ct$, whereas for the diffusion equation the T equation $T' + \kappa^2\alpha^2 T = 0$ gives the exponential decay $\exp(-\kappa^2\alpha^2 t)$.

Finally, note that in this chapter on the wave equation there is no section on numerical solution analogous to Section 18.6 on the numerical solution of the diffusion equation by the method of finite differences. The finite-difference method can indeed be applied to the wave equation, but if the initial conditions $y(x, 0) = f(x)$ and $y_t(x, 0) = g(x)$ are not smooth functions, then inaccuracies result from the discontinuities that propagate into the solution domain. In that case it is better to use the **method of characteristics**, which uses the more natural characteristic variables ξ and η in place of x and t. Essentially, the calculation is carried out along the characteristics.*

*For an introduction to the method of characteristics see G. D. Smith, *Numerical Solution of Partial Differential Equations* (New York: Oxford University Press, 1965) or, for a more complete discussion, see E. Zauderer, *Partial Differential Equations of Applied Mathematics* (New York: Wiley, 1983).

Chapter 20

Laplace Equation

20.1 Introduction

We have already encountered the **Laplace equation**

$$\boxed{\nabla^2 u = 0} \tag{1}$$

in Chapter 16, as well as its nonhomogeneous version, the **Poisson equation**

$$\boxed{\nabla^2 u = f,} \tag{2}$$

where f is a "source" function that is prescribed over the region in question. Recall, for example, the unsteady diffusion equation

$$\alpha^2 \nabla^2 u = u_t - F(x, y, z, t) \tag{3}$$

governing the temperature field $u(x, y, z, t)$, in which $F(x, y, z, t)$ is a heat source distribution. If F does not vary with t and if there exists a steady-state solution $u(x, y, z)$, then the latter satisfies the Poisson equation $\nabla^2 u = -F(x, y, z)/\alpha^2$. If there is no heat source distribution [i.e., if $F(x, y, z) = 0$], then the steady-state temperature distribution $u(x, y, z)$ satisfies the Laplace equation (1).

As a second example observe that the electric potential (i.e., the voltage) field $\Phi(x, y, z)$ is governed by the Poisson equation

$$\nabla^2 \Phi = -\frac{1}{\epsilon}\, q(x, y, z), \tag{4}$$

where $q(x, y, z)$ is the charge density distribution (which serves as a "source" for the electric potential) and ϵ is a physical constant known as the permittivity of the medium. If $q(x, y, z) = 0$ in the region, then Φ satisfies the Laplace equation $\nabla^2 \Phi = 0$. Since the presence of an electric field is solely attributable to the presence of charges, how can there be anything other than the trivial solution $\Phi(x, y, z) = 0$ in the event that $q(x, y, z) = 0$? The answer is that there may be charges *outside of* the region under consideration or on its boundary.

Finally, recall from Example 3 of Section 16.10 that the velocity potential $\Phi(x, y, z)$ for any irrotational incompressible fluid flow is governed by the Laplace equation

$$\nabla^2 \Phi = 0,$$

where Φ is related to the velocity field $\mathbf{v}(x, y, z)$ by the formula $\mathbf{v} = \nabla \Phi$.

Emphasis in this chapter is on the Laplace equation, with the Poisson equation considered only within the exercises. Like Chapters 18 and 19, Chapter 20 is organized according to the various methods of solution. In Sections 20.2 and 20.3 we study the solution of the Laplace equation by separation of variables – using Cartesian coordinates in Section 20.2 and non-Cartesian coordinates in Section 20.3. Solution of certain problems by the Fourier transform is covered in Section 20.4, and the numerical finite difference method is the subject of Section 20.5.

20.2 Separation of Variables; Cartesian Coordinates

We limit our attention in this chapter to *two*-dimensional problems, so the domain $\mathcal{D}$ is some part of the x, y plane. For the method of separation of variables to work, $\mathcal{D}$ must be bounded by constant-coordinate curves, so if we use the Cartesian coordinates x, y, then the generic domain is a rectangle, bounded by constant-x and constant-y lines.

EXAMPLE 1. *Dirichlet Problem for Rectangle.* Consider the boundary-value problem

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \quad \text{in } \mathcal{D}, \tag{1a}$$

$$u(0, y) = 0, \quad (0 < y < b) \tag{1b}$$

$$u(a, y) = f(y), \quad (0 < y < b) \tag{1c}$$

$$u(x, 0) = u(x, b) = 0, \quad (0 < x < a) \tag{1d}$$

where $\mathcal{D}$ is the rectangle shown in Fig. 1. Since all of the boundary conditions are of Dirichlet type (i.e., where u is given), we call (1) a **Dirichlet problem**.

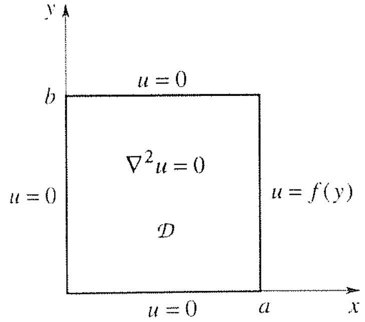


Figure 1. The Dirichlet problem (1).

To solve by separation of variables, seek

$$u(x, y) = X(x)Y(y). \tag{2}$$

Putting (2) into (1a) and separating the variables gives

$$\frac{X''}{X} = -\frac{Y''}{Y} = \text{constant} = \kappa^2, \tag{3}$$

so

$$X'' - \kappa^2 X = 0, \tag{4a}$$

$$Y'' + \kappa^2 Y = 0, \tag{4b}$$

and

$$X(x) = \begin{cases} A + Bx, & \kappa = 0 \\ C\cosh \kappa x + D\sinh \kappa x, & \kappa \neq 0 \end{cases} \tag{5a}$$

$$Y(y) = \begin{cases} E + Fy, & \kappa = 0 \\ G\cos \kappa y + H\sin \kappa y, & \kappa \neq 0 \end{cases} \tag{5b}$$

Why did we choose the constant in (3) as $+\kappa^2$ rather than as $-\kappa^2$? Because, looking ahead to the application of the boundary conditions we anticipate the eventual Fourier series expansion of $f(y)$. The choice $+\kappa^2$ does indeed lead to the $\cos \kappa y$ and $\sin \kappa y$ solutions in (5b) that will be needed for that Fourier series expansion; $-\kappa^2$ would have led to $\cosh \kappa y$ and $\sinh \kappa y$ instead.

Because the Laplace equation $\nabla^2 u = 0$ is linear, we can use superposition and combine the $\kappa = 0$ and $\kappa \neq 0$ product solutions as

$$\begin{aligned} u(x,y) &= (A + Bx)(E + Fy) \\ &\quad +(C\cosh \kappa x + D\sinh \kappa x)(G\cos \kappa y + H\sin \kappa y). \end{aligned} \tag{6}$$

Saving the boundary condition at $x = a$ for last, we first apply the boundary conditions on the edges $y = 0$, $y = b$, and $x = 0$:

$$u(x,0) = 0 = (A + Bx)E + (C\cosh \kappa x + D\sinh \kappa x)G,$$

so (to retain as robust a solution as possible) set $E = 0$ (rather than $A = 0$ and $B = 0$) and $G = 0$ (rather than $C = 0$ and $D = 0$). Then (6) becomes

$$u(x,y) = (I + Jx)y + (P\cosh \kappa x + Q\sinh \kappa x)\sin \kappa y, \tag{7}$$

where we have combined AF as I, BF as J, CH as P, and DH as Q for brevity. Next,

$$u(x,b) = 0 = (I + Jx)b + (P\cosh \kappa x + Q\sinh \kappa x)\sin \kappa b \tag{8}$$

gives $I = 0$, $J = 0$, and $\sin \kappa b = 0$. Hence,

$$\kappa = n\pi/b \quad (n = 1, 2, \dots)$$

so, with the help of superposition, we can update (7) as

$$u(x,y) = \sum_{n=1}^{\infty} \left(P_n \cosh \frac{n\pi x}{b} + Q_n \sinh \frac{n\pi x}{b}\right) \sin \frac{n\pi y}{b}. \tag{9}$$

Next, the "western" boundary condition (i.e., at $x = 0$) gives

$$u(0,y) = 0 = \sum_{n=1}^{\infty} P_n \sin \frac{n\pi y}{b}, \qquad (0 < y < b) \tag{10}$$

which is satisfied by setting $P_n = 0$ for $n = 1, 2, \dots$. Thus, (9) becomes

$$u(x,y) = \sum_{n=1}^{\infty} Q_n \sinh \frac{n\pi x}{b} \sin \frac{n\pi y}{b}. \tag{11}$$

Finally, the eastern boundary condition

$$u(a,y) = f(y) = \sum_{n=1}^{\infty} Q_n \sinh \frac{n\pi a}{b} \sin \frac{n\pi y}{b} \qquad (0 < y < b) \tag{12}$$

is seen to be a half-range sine series so

$$Q_n \sinh \frac{n\pi a}{b} = \frac{2}{b}\int_0^b f(y) \sin \frac{n\pi y}{b}\, dy,$$

or

$$Q_n = \frac{2}{b \sinh \dfrac{n\pi a}{b}} \int_0^b f(y) \sin \frac{n\pi y}{b}\, dy. \tag{13}$$

The solution to (1) is given by (11) and (13).

COMMENT 1. We stated that (10) is satisfied by setting $P_n = 0$ for $n = 1, 2, \dots$. Obviously that is true, but there is a subtle question here: *must* the P_n's be zero? That is, might the $\sin(n\pi y/b)$ terms cancel to zero on $0 < y < b$ without all the P_n's being zero? Surely the $\sin(n\pi y/b)$ terms are linearly independent on $0 < y < b$, and if a *finite* linear combination of linearly independent functions is zero, then each coefficient must be zero. However, the sum in (10) is an infinite series, not a linear combination of a finite number of terms. The cleanest way to handle (10) is to see that it is actually a half-range sine expansion of the function $u(0, y) = 0$. Thus,

$$P_n = \frac{2}{b}\int_0^b (0) \sin \frac{n\pi y}{b}\, dy = 0,$$

as stated.

COMMENT 2. To make our results more concrete, consider a specific $f(y)$. For simplicity, let $f(y)$ be a constant, say

$$f(y) = 100.$$

Then (13) gives

$$Q_n = \frac{400}{n\pi \sinh \dfrac{n\pi a}{b}}$$

for $n = 1, 3, \dots$ and 0 for $n = 2, 4, \dots$, so (11) becomes

$$u(x,y) = \frac{400}{\pi} \sum_{n=1,3,\dots}^{\infty} \frac{1}{n} \frac{\sinh(n\pi x/b)}{\sinh(n\pi a/b)} \sin \frac{n\pi y}{b}. \tag{14}$$

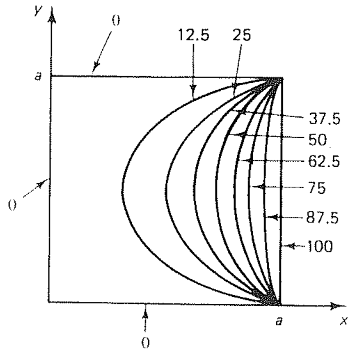


Figure 2. Selected isotherms for $f(y) = 100$ and $b = a$.

One useful way of presenting such two-dimensional results is to plot a number of $u =$ constant curves, isothermal curves if we consider (1) in the context of steady-state heat conduction. We have done so in Fig. 2 for the case where $b = a$. The $u = 0$ isotherm is the northern, western, and southern boundary; the $u = 100$ isotherm is the eastern edge; and all other isotherms spring from the corners $(a, 0)$ and (a, a). As implied by Fig. 2, all u values within the rectangle are between the minimum and maximum values of u on the boundary, namely, 0 and 100, respectively. This result illustrates the important and fundamental **maximum principle** of potential theory (i.e., the theory associated with the Laplace equation), which states that if the Laplace equation $\nabla^2 u = 0$ holds on a domain $\mathcal{D}$, then the maximum (and minimum) values of u occur on the boundary of $\mathcal{D}$, not in its interior. This result is proved in the next section.

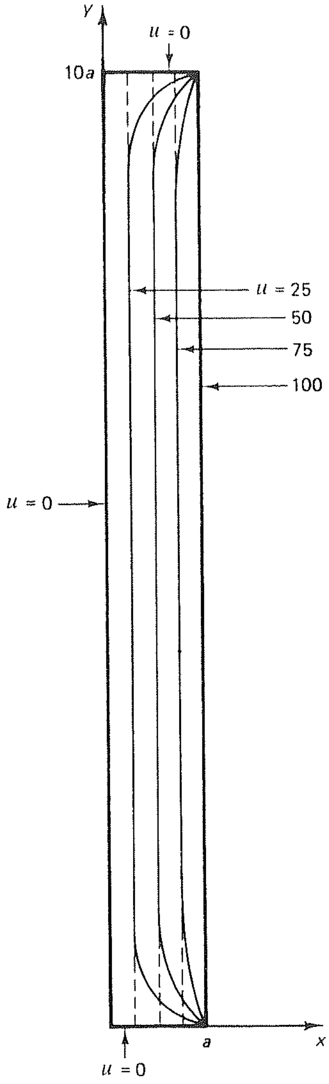


Figure 3. High-aspect-ratio case.

COMMENT 3. We have also plotted the isotherms corresponding to (14) for the high-aspect-ratio case where $b = 10a$, again with $f(y) = 100$ (Fig. 3). Observe that over $2a < y < 8a$, say, the problem is essentially one-dimensional, with u varying with x but hardly at all with y. If, accordingly, we neglect the $\partial^2 u/\partial y^2$ term in the Laplace equation, then we have the problem

$$\frac{\partial^2 u}{\partial x^2} \approx 0; \qquad u\Big|_{x=0} = 0, \quad u\Big|_{x=a} = 100, \tag{15}$$

with solution

$$u(x,y) \approx 100\frac{x}{a}, \tag{16}$$

the isotherms of which are the lines x = constant in Fig. 3. We see that (16) is an excellent approximation to $u(x,y)$ except near the ends, that is, except within one or two widths (the width being a) of the ends $y = 0$ and $y = 10a$. Indeed, if we are interested in the solution only within $2a < y < 8a$, say, then the simpler one-dimensional model and its solution (16) may well suffice in place of the two-dimensional model and the more cumbersome solution (14). Of course, (16) does not satisfy the boundary conditions $u(x,0) = 0$ and $u(x,10a) = 0$, so there are regions of adjustment near those two ends, where the approximate solution (16) needs to be blended with the boundary conditions $u(x,0) = 0$ and $u(x,10a) = 0$ in a way that satisfies the Laplace equation.

The results found in this example hold in general; namely, we can expect end effects to be significant only within one or two widths of the end. ▮

Remember that in Chapters 18 and 19 we always chose the separation constant to be $-\kappa^2$, and that we always applied the boundary conditions before the initial condition. For the Laplace equation, however, the sign of the separation constant and the sequencing of the boundary conditions needs to be decided on a case-by-case basis. We offer this rule of thumb as guidance:

Anticipating the edge along which the eventual Fourier series will take place, choose the $+\kappa^2$ or $-\kappa^2$ so as to obtain oscillatory functions along that edge. Then, apply the boundary conditions adjacent to that edge first.

For instance, in Example 1 we can anticipate that it is the boundary condition $u(a,y) = f(y)$ that will require the Fourier series, so we choose $+\kappa^2$ in (3) so as to obtain the oscillatory solutions $Y(y) = \cos\kappa y$ and $\sin\kappa y$ in the y variable. Then, we apply the boundary conditions on the northern and southern edges, which are adjacent to that edge. Similarly in Chapters 18 and 19. There, the Fourier series is always along the southern edge in the x, t plane (i.e., along the edge $t = 0$) so we chose $-\kappa^2$ in order to obtain $\cos\kappa x$ and $\sin\kappa x$; we applied the boundary conditions on the adjacent edges $x = 0$ and $x = L$ first, and on the edge $t = 0$ last.

What if the boundary conditions are, in place of those in Fig. 1, $u(0,y) = p(y)$, $u(a,y) = f(y)$, $u(x,0) = q(x)$, and $u(x,b) = g(x)$? Evidently *each* edge will require a Fourier series expansion, so how can we apply the adjacent boundary conditions first if they require Fourier series expansions themselves? This difficulty can be resolved by using the concepts of linearity and superposition to break the problem into four simpler ones as indicated schematically in Fig. 4. The idea is to solve each of the four problems on the right along the lines indicated in Example 1 (using $+\kappa^2$ for the first and third problems and $-\kappa^2$ for the second and fourth) and then obtain u as the sum

$$u(x, y) = u_1(x, y) + u_2(x, y) + u_3(x, y) + u_4(x, y). \tag{17}$$

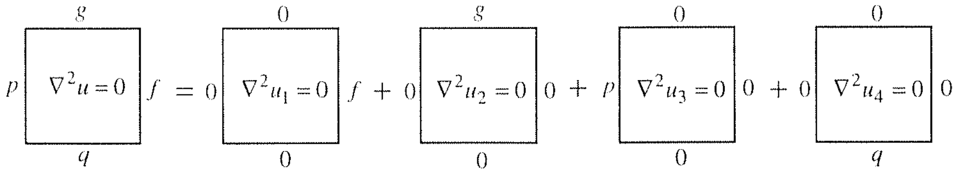


Figure 4. Use of superposition.

To see that (17) is true, add the equations $\nabla^2 u_1 = 0$, $\nabla^2 u_2 = 0$, $\nabla^2 u_3 = 0$, and $\nabla^2 u_4 = 0$, and obtain

$$\nabla^2 u_1 + \nabla^2 u_2 + \nabla^2 u_3 + \nabla^2 u_4 = 0. \tag{18}$$

But since the operator ∇^2 is linear, it follows from (18) that

$$\nabla^2 (u_1 + u_2 + u_3 + u_4) = 0,$$

so $u = u_1 + u_2 + u_3 + u_4$ does satisfy the Laplace equation, as required. Turning to the boundary conditions, consider the eastern condition. From (17) and the conditions imposed on u_1, u_2, u_3, and u_4,

$$\begin{aligned} u(a, y) &= u_1(a, y) + u_2(a, y) + u_3(a, y) + u_4(a, y) \\ &= f(y) + 0 + 0 + 0 \\ &= f(y), \end{aligned}$$

as required. Similarly for the other boundary conditions.

EXAMPLE 2. *Semi-Infinite Strip.* Consider the Dirichlet problem

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \quad \text{in } \mathcal{D}, \tag{19a}$$
$$u(0, y) = 20, \quad u(5, y) = 50, \quad (0 < y < \infty) \tag{19b}$$
$$u(x, 0) = f(x), \quad (0 < x < 5) \tag{19c}$$
$$u(x, y) \quad \text{bounded as} \quad y \to \infty, \tag{19d}$$

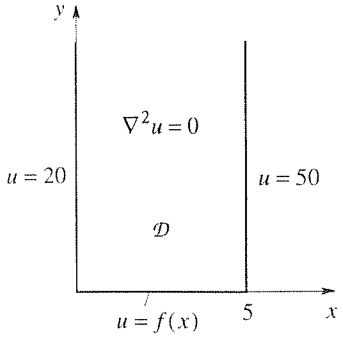


Figure 5. Semi-infinite strip.

where $\mathcal{D}$ is the semi-infinite strip $0 < x < 5$, $0 < y < \infty$ (Fig. 5).

Since physical objects are of finite size, why might we be interested in a semi-infinite strip that extends to $y = \infty$? The actual physical object might, for instance, be a slender "fin" attached to a thick base as shown in Fig. 6. Then the problem is actually defined on a complicated three-dimensional domain. However, suppose that our interest is not in finding the temperature field everywhere in the object but only near the end of the fin, within the rectangle $ABCE$. And suppose that we know the boundary conditions $u = 20$ along AE, $u = 50$ along BC, and $u = f(x)$ along EC. We could take our domain $\mathcal{D}$ to be the rectangle $ABCE$, but we do not know a boundary condition along AB. To within a good approximation, we render the problem tractable by letting $\mathcal{D}$ be the entire semi-infinite strip shown in Fig. 5, with the boundedness condition (19d) as our missing boundary condition at $y = \infty$. Based on Comment 3 of Example 1, we expect the difference between the actual temperature field and the one defined by (19) to be very small within the region of interest, $ABCE$.

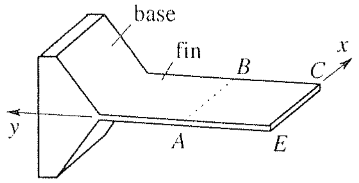


Figure 6. The actual object.

We anticipate that the eventual Fourier series expansion will be a half- or quarter-range expansion on the edge $y = 0$. Thus, to obtain oscillatory functions of x (namely $\cos \kappa x$ and $\sin \kappa x$) rather than of y, write

$$\frac{X''}{X} = -\frac{Y''}{Y} = -\kappa^2 \tag{20}$$

in place of (3). Solving the resulting ODE's on X and Y and superimposing the $\kappa = 0$ and $\kappa \neq 0$ solutions gives

$$\begin{aligned} u(x,y) &= (A + Bx)(E + Fy) \\ &\quad + (C \cos \kappa x + D \sin \kappa x)(Ge^{\kappa y} + He^{-\kappa y}). \end{aligned} \tag{21}$$

Apply the boundedness condition (19d) first. Since the y and $e^{\kappa y}$ terms in (21) grow unboundedly as $y \to \infty$, we must set $F = 0$ and $G = 0$ to eliminate those terms. Then (21) becomes

$$u(x,y) = I + Jx + (P \cos \kappa x + Q \sin \kappa x)e^{-\kappa y}, \tag{22}$$

where we have combined AE as I, BE as J, CH as P, and DH as Q. Since we anticipate the Fourier expansion to be on the $y = 0$ edge, we save that boundary condition for last. Next,

$$u(0,y) = 20 = I + Pe^{-\kappa y}, \tag{23}$$

and matching the coefficients of the (linearly independent) constant and $e^{-\kappa y}$ terms on both sides of (23) gives $20 = I$ and $0 = P$. Using these results to update (22) gives

$$u(x,y) = 20 + Jx + Q \sin \kappa x \, e^{-\kappa y}. \tag{24}$$

Next,

$$u(5,y) = 50 = 20 + 5J + Q \sin 5\kappa \, e^{-\kappa y}, \tag{25}$$

so $50 = 20 + 5J$ and $\sin 5\kappa = 0$. Thus, $J = 6$ and $\kappa = n\pi/5$ ($n = 1, 2, \ldots$). Putting these results into (24) we have, with the help of superposition,

$$u(x,y) = 20 + 6x + \sum_{n=1}^{\infty} Q_n \sin \frac{n\pi x}{5} e^{-n\pi y/5}. \tag{26}$$

Finally,

$$u(x,0) = f(x) = 20 + 6x + \sum_{n=1}^{\infty} Q_n \sin \frac{n\pi x}{5} \tag{27}$$

or, moving the (known) $20 + 6x$ terms to the left-hand side,

$$f(x) - 20 - 6x = \sum_{n=1}^{\infty} Q_n \sin\frac{n\pi x}{5}. \qquad (0 < x < 5) \tag{28}$$

The latter is a half-range sine expansion of $f(x) - 20 - 6x$, so we can compute the Q_n's from

$$Q_n = \frac{2}{5}\int_0^5 [f(x) - 20 - 6x]\sin\frac{n\pi x}{5}\,dx. \tag{29}$$

Thus, the solution to (19) is given by (26), with the Q_n's computed according to (29).

COMMENT 1. The $\kappa = 0$ term $(A + Bx)(E + Fy)$ in (21) contributes the $20 + 6x$ part of the final solution (26). The graph of that part of the solution is like a ramp, from $u = 20$ along the $x = 0$ edge to $u = 50$ along the $x = 5$ edge. In the language of Chapter 18 we can think of $20 + 6x$ as the "steady-state" solution, the solution that is approached as $y \to \infty$ (analogous to the limit $t \to \infty$ in Chapter 18), and we can think of the series in (26) as the "transient" that blends the "steady-state" $20 + 6x$ with the "initial condition" $u(x, 0) = f(x)$. We see from (26) that the transient part, or *end effect*, dies out exponentially with y; with $n = 1$, the $\exp(-n\pi y/5)$ factor is merely 0.043 at $y = 5$ and 0.0019 at $y = 10$, and with $n = 2, 3, \ldots$ it is even smaller. Only if $f(x)$ happens to equal $20 + 6x$ does the end effect vanish entirely, for then all the Q_n's are zero.

COMMENT 2. In (21) we wrote $Ge^{\kappa y} + He^{-\kappa y}$ but could have written $R\cosh\kappa y + S\sinh\kappa y$, say, instead. The choice is immaterial since the two forms are equivalent, but the former is more convenient regarding the application of the boundedness condition, for it is clear that $e^{\kappa y}$ is a "bad" term (unbounded) and that $e^{-\kappa y}$ is a "good" term (bounded) so we set $G = 0$. Working with $\cosh\kappa y$ and $\sinh\kappa y$ instead would be awkward because *both* are unbounded (Fig. 7). However,

$$\begin{aligned} R\cosh\kappa y + S\sinh\kappa y &= \frac{R}{2}\left(e^{\kappa y} + e^{-\kappa y}\right) + \frac{S}{2}\left(e^{\kappa y} - e^{-\kappa y}\right) \\ &= \frac{R+S}{2}e^{\kappa y} + \frac{R-S}{2}e^{-\kappa y}, \end{aligned} \tag{30}$$

so we can arrange for the unbounded parts of the $\cosh\kappa y$ and $\sinh\kappa y$ to cancel by choosing $S = -R$, in which case (30) reduces to $Re^{-\kappa y}$. Thus, we arrive at the same place but the trip is more arduous.

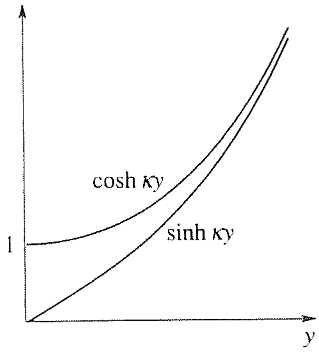


Figure 7. $\cosh\kappa y$ and $\sinh\kappa y$.

COMMENT 3. Since we had three nonzero boundary conditions (Fig. 5), why did we not break the problem into three sub-problems along the lines indicated in Fig. 4? We could have but did not need to because the boundary conditions along the edges $x = 0$ and $x = 5$ are merely constants and can therefore be handled by the $(A + Bx)(E + Fx)$ *ramp term* in the solution.

COMMENT 4. In these first two examples the Fourier expansions happened to be half-range sine series, but that will not always be the case and will depend on the boundary conditions. For instance, if we change the Dirichlet boundary condition $u(0, y) = 20$ to a Neumann boundary condition such as $u_x(0, y) = 3$, then in place of the half-range sine expansion (28) we would have a quarter-range cosine expansion. ▮

Closure. In this section we study the separation-of-variable solution of the two-dimensional Laplace equation $u_{xx} + u_{yy} = 0$ on domains bounded by constant-x and constant-y lines. Unlike Chapters 18 and 19, where we always take the separation constant to be $-\kappa^2$ and apply the boundary conditions at $x = 0$ and $x = L$ before the initial condition(s) at $t = 0$, here we need to choose the sign of κ^2 and the sequence of application of the boundary conditions on a case-by-case basis. The rule of thumb is to choose the separation constant as $+\kappa^2$ or $-\kappa^2$ so as to give oscillatory solutions (cosines and sines) along the edge where the eventual Fourier expansion will take place, and then to be sure to apply the boundary conditions adjacent to that edge first. (For example, if the Fourier expansion is on the eastern edge, then by the "adjacent" edges we mean the northern and southern ones.)

Consider the general Dirichlet problem on u shown in Fig. 4. If all of the functions $p(y), g(x), f(y)$, and $q(x)$ are nonconstant, then we can break the problem into four sub-problems, as shown in the figure, and solve separately for $u_1(x, y)$, $u_2(x, y)$, $u_3(x, y)$, and $u_4(x, y)$. For the u_1 and u_3 problems, set $X''/X = -Y''/Y = +\kappa^2$, so as to obtain cosine and sine solutions in the y variable, and apply the southern and northern boundary conditions before attempting the nonhomogeneous boundary condition (eastern in the u_1 problem, western in the u_3 problem). In fact, you will find that if the boundary conditions on any two opposite boundaries are constants, then it is not necessary to break the problem down as we do in Fig. 4 (although we can if we wish). For instance, suppose that both $p(y)$ and $f(y)$ are constants. Then set $X''/X = -Y''/Y = -\kappa^2$ to obtain cosine and sine solutions in the x variable, and apply the western and eastern boundary conditions first.

Similar statements apply if boundary conditions of Neumann or Robin type are present on one or more edges, but be careful because if all four edges have Neumann boundary conditions, then there may be no solution or a nonunique solution (Exercises 17 and 18).

In Example 2 we consider the domain to be an idealized semi-infinite strip and adopt a boundedness condition in lieu of the missing boundary condition at $y = \infty$. The Fourier expansion is necessarily on the finite edge (the edge $y = 0$ in Example 2). Given the choice between expressing Y in terms of $\cosh \kappa y$ and $\sinh \kappa y$ or in terms of $e^{\kappa y}$ and $e^{-\kappa y}$, we choose the latter because of convenience in regard to the application of the boundedness condition.

Finally, and very important, is the fact that the Laplace and Poisson equations arise in the context of **boundary-value** problems; that is, a boundary condition is supplied on each of the four edges. In contrast, the diffusion and wave problems studied in Chapters 18 and 19 are of initial-value type (with respect to the t variable) since conditions are given at $t = 0$ but not at a final time or at $t = \infty$. This boundary-value nature will be felt more acutely when we use the finite-difference method in Section 20.5.

EXERCISES 20.2

1. Solve $u_{xx}+u_{yy}=0$ in the rectangle $0<x<3, 0<y<2$ by separation of variables, subject to the given boundary conditions. H denotes the Heaviside function.

(a) $u(0,y)=u(x,2)=u(3,y)=0,\ u(x,0)=50\sin(\pi x/3)$
(b) $u(0,y)=u(x,0)=u(3,y)=0,\ u(x,2)=10\sin(\pi x/3)-4\sin\pi x$
(c) $u(x,0)=u(3,y)=u(x,2)=0,\ u(0,y)=5\sin\pi y+4\sin 2\pi y-\sin 3\pi y$
(d) $u(0,y)=u(x,2)=u(3,y)=0,\ u(x,0)=50H(x-2)$
(e) $u_x(0,y)=u(x,2)=u(3,y)=0,\ u(x,0)=50H(x-2)$
(f) $u(0,y)=u(x,2)=u_x(3,y)=0,\ u(x,0)=50H(x-2)$
(g) $u_x(0,y)=u(x,2)=u_x(3,y)=0,\ u(x,0)=50H(x-2)$
(h) $u_y(x,2)=u(3,y)=u(x,0)=0,\ u(0,y)=H(y-1)$
(i) $u(0,y)=u(x,0)=u(3,y)=0,\ u(x,2)=5x$
(j) $u(0,y)=u(x,2)=u(3,y)=0,\ u(x,0)=5[H(x-1)-H(x-2)]$
(k) $u(x,2)=u(3,y)=u(x,0)=0,\ u_x(0,y)=5\sin 3\pi y$
(l) $u(x,2)=u(3,y)=u(x,0)=0,\ u_x(0,y)=20$
(m) $u(x,2)=u(3,y)=u_y(x,0)=0,\ u_x(0,y)=20$
(n) $u_y(x,2)=u(3,y)=u_y(x,0)=0,\ u_x(0,y)=20$

2. (a) The solution to Exercise 1(d) is given in the Answers to Selected Exercises. Using that solution and computer software, evaluate $u(2.5,1)$, $u(2.5,0.5)$, $u(2.5,0.2)$, and $u(2.5,0.1)$, correct to two decimal places. In each case, tell how many terms must be summed to achieve that accuracy. Explain why more terms are needed as the point approaches the x axis.
(b) The same as part (a), but using the solution to Exercise 1(h).
(c) In Exercise 1(e) evaluate $u(1,1)$ to three significant figures.
(d) In Exercise 1(e) evaluate $u(0,y)$ at $y=0.25,0.5,0.75,\dots,1.75$, to three significant figures, and plot $u(0,y)$ versus y, by hand or by computer.

3. Solve $u_{xx}+u_{yy}=0$ in the square $0<x<2$, $0<y<2$ by separation of variables, subject to the given boundary conditions. (You should be able to obtain the solution in closed form.) Then, obtain a computer plot of the $u=10,20,30,\dots,90$ isotherms using software such as the *Maple* implicitplot command. NOTE: We urge you to try sketching the isotherms even before you solve the problem.

(a) $u(0,y)=u(x,2)=u(2,y)=0,\ u(x,0)=100\sin(\pi x/2)$
(b) $u(0,y)=u(2,y)=0,\ u(x,2)=u(x,0)=100\sin(\pi x/2)$
(c) $u(0,y)=u(x,2)=0,\ u(2,y)=100\sin(\pi y/2),\ u(x,0)=100\sin(\pi x/2)$
(d) $u(0,y)=u(x,2)=0,\ u_x(2,y)=0,\ u(x,0)=100\sin(\pi x/4)$
(e) $u(0,y)=u(x,2)=0,\ u(x,0)=100\sin(\pi x/2),\ u(2,y)=20\sin(\pi y/2)$
(f) $u(0,y)=u(2,y)=0,\ u(x,0)=100\sin(\pi x/2),\ u(x,2)=20\sin(\pi x/2)$
(g) $u(0,y)=u(2,y)=100\sin(\pi y/2),\ u(x,0)=u(x,2)=100\sin(\pi x/2)$

4. Solve $u_{xx}+u_{yy}=0$ in the rectangle $0<x<a$, $0<y<b$ subject to the boundary conditions $u(0,y)=p(y)$, $u(x,b)=u_2$, $u(a,y)=f(y)$, $u(x,0)=u_1$ *without* breaking the problem into sub-problems; $p(y)$ and $f(y)$ are prescribed functions and u_1 and u_2 are prescribed constants. HINT: Read the second paragraph of the closure.

5. Same as Exercise 4, but with these boundary conditions:

(a) $u(0,y)=u_1,\ u(x,b)=p(x),\ u(a,y)=u_2,\ u(x,0)=f(x)$
(b) $u_x(0,y)=p(y),\ u(x,b)=u_1,\ u(a,y)=f(y),\ u(x,0)=u_2$
(c) $u_x(0,y)=p(y),\ u(x,b)=u_1,\ u_x(a,y)=f(y),\ u(x,0)=u_2$

6. Solve $u_{xx}+u_{yy}=0$ in the rectangle $0<x<a, 0<y<b$ subject to the boundary conditions $u(x,0)=u_1$, $u(x,b)=u_2$, $u(0,y)=u_3$, $u(a,y)=u_4$, where $u_1,\dots,u_4$ are constants. Do not break the problem into subproblems; you don't need to. Rather, choose $X''/X=-Y''/Y=+\kappa^2$ and apply the southern and northern boundary conditions first. (You may leave expansion coefficients in integral form.) Next, solve the problem again, this time choosing $X''/X=-Y''/Y=-\kappa^2$ and applying the western and eastern boundary conditions first. Your two solutions will look different but will merely be two different representations of the same function. If $b=10a$, which of the two solution forms would you prefer – for purposes of calculation? Explain your reasoning.

7. Let $f(y)=100$ in (1c), and let $b=a$. Without solving the problem (or using the solution in the text), show that $u(a/2,a/2)=25$. HINT: Let $p(y)=g(x)=f(y)=q(x)=100$ in Fig. 4.

8. To promote physical "intuition," we ask you to draw a neat, labeled sketch of representative isotherms for the problem consisting of the Laplace equation on the square $0 < x < a$, $0 < y < a$ with the given boundary conditions.

(a) $u(0, y) = u(x, a) = u(a, y) = 0$, $u(x, 0) = 100 \sin(\pi x/a)$
(b) $u(0, y) = u(x, a) = u_x(a, y) = 0$, $u(x, 0) = 100 \sin(\pi x/2a)$
(c) $u(0, y) = u(x, 0) = 0$, $u(a, y) = u(x, a) = 20$
(d) $u(0, y) = u(x, 0) = u(a, y) = 0$, $u(x, a) = 100$ for $0 < x < a/2$ and 0 for $a/2 < x < a$
(e) $u(0, y) = u(x, 0) = 0$, $u(x, a) = 0$, $u(a, y) = 20$

9. Show that the solution (11) of the problem (1) can be expressed in the form of an integration over the boundary data, namely,

$$u(x, y) = \int_0^b K(\eta; x, y) f(\eta)\, d\eta, \tag{9.1}$$

and give an expression for the kernel $K(\eta; x, y)$; it will be in the form of an infinite series.

10. Solve $u_{xx} + u_{yy} = 0$ in the semi-infinite strip $0 < x < \infty$, $0 < y < 1$ subject to the given boundary conditions plus the condition that u is bounded as $x \to \infty$.

(a) $u(0, y) = 0$, $u(x, 0) = 10$, $u_y(x, 1) = 0$
(b) $u(0, y) = 100$, $u_y(x, 0) = u_y(x, 1) = 0$
(c) $u_x(0, y) = 5$, $u(x, 0) = u(x, 1) = 0$
(d) $u(0, y) = 0$, $u(x, 0) = 50$, $u(x, 1) = 10$
(e) $u(0, y) = 10y$, $u(x, 0) = 20$, $u(x, 1) = 50$

11. (a)–(e) Give a labeled sketch of representative isotherms for the corresponding problem in Exercise 10.

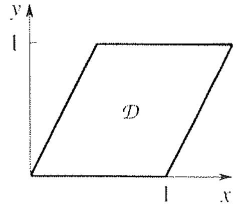
12. The problem

$$u_{xx} + u_{yy} = 0,$$
$$u(x, 0) = 0, \quad u(x, b) = 50e^{-(x/10b)^2}$$

on $-\infty < x < \infty$, $0 < y < b$ admits a simple and accurate approximate solution, which we ask you to find. HINT: $\exp\left[-(x/10b)^2\right]$ is a slowly varying function of x.

13. In Example 2 we apply a boundedness condition on u at $y = \infty$. Dropping that condition, put forward two or three solutions that are unbounded on the semi-infinite strip.

14. Consider the Laplace equation $u_{xx} + u_{yy} = 0$ on the parallelogram $\mathcal{D}$ shown, bounded by the lines $y = 0$, $y = 1$, $y = 2x$, and $y = 2x - 2$, with boundary conditions given on the four edges. There is no future in seeking $u(x, y) = X(x)Y(y)$ and using separation of variables because the boundary is not comprised of constant coordinate curves. Specifically, the left and right edges are neither constant-x nor constant-y lines. One possibility seems to be a change of variables from x, y to ξ, η according to



$$\xi = y - 2x, \quad \eta = y \tag{14.1}$$

so that the new domain, in the ξ, η plane, will be a rectangle bounded by constant-ξ and constant-η lines.

(a) Show that new domain in a labeled sketch.
(b) Show that in terms of ξ and η the Laplace equation becomes

$$5u_{\xi\xi} + 2u_{\xi\eta} + u_{\eta\eta} = 0 \tag{14.2}$$

and that our plan fails because (14.2) is not separable. NOTE: Nonetheless, the idea is a good one; we simply need to figure out how to design a change of variables $\xi = F(x, y)$ so as to simplify the domain without at the same time complicating the PDE. How to do this, for the two-dimensional Laplace equation, is the subject of Chapter 22 on conformal mapping.

15. (*Poisson equation*) Consider the Poisson problem

$$\begin{aligned} u_{xx} + u_{yy} &= f(x, y), \\ u(0, y) = u(x, b) = u(a, y) &= u(x, 0) = 0 \end{aligned} \tag{15.1}$$

on the rectangle $0 < x < a$, $0 < y < b$, where the source function $f(x, y)$ is prescribed.

(a) Solve (15.1) by separation of variables for the case where $f(x, y) = \text{constant} \equiv f$, leaving expansion coefficients in integral form. HINT: Noticing that $fx^2/2$ is a simple particular solution of (15.1a), seek u in the form

$$u(x, y) = \frac{f}{2} x^2 + U(x, y). \tag{15.2}$$

Show that the "homogeneous solution" U satisfies the problem

$$\begin{aligned} U_{xx} + U_{yy} &= 0, \\ U(0, y) = 0, \quad U(a, y) &= -fa^2/2, \\ U(x, 0) = U(x, b) &= -fx^2/2, \end{aligned} \tag{15.3a,b,c}$$

and solve for U by separation of variables. NOTE: $fx^2/2$ is not the only particular solution that could be used; for instance, we could use $fy^2/2$, $fx^2/2 + 37xy - 5y + 6$, and so on, but $fx^2/2$ (or $fy^2/2$) seems a simple and natural choice.
(b) Observe that the method proposed in (a) will work not only when $f(x, y)$ is a constant, but also when it is a function only of x or only of y. Here, we ask you to solve (15.1) for the case where $f(x, y)$ is not necessarily of that form. HINT: Use the **eigenvector expansion method** in very much the same manner as we did in Exercise 17 of Section 18.3. Essentially, we can use either the eigenfunctions $\sin n\pi x/a$ in the x variable or the eigenfunctions $\sin n\pi y/a$ in the y variable to expand "everything in sight." Specifically, expanding

$$f(x, y) = \sum_{n=1}^{\infty} f_n(x) \sin \frac{n\pi y}{b}, \tag{15.4}$$

where

$$f_n(x) = \frac{2}{b} \int_0^b f(x, y) \sin \frac{n\pi y}{b}\, dy, \tag{15.5}$$

and seeking

$$u(x, y) = \sum_{n=1}^{\infty} g_n(x) \sin \frac{n\pi y}{b}, \tag{15.6}$$

show that the g_n's are found by solving the problems

$$g_n'' - \frac{n^2\pi^2}{b^2} g_n = f_n(x); \qquad g_n(0) = g_n(a) = 0 \tag{15.7}$$

for $n = 1, 2, \ldots$.
(c) Implement the method of part (b) for the case where $f(x, y) = xy$, and solve for $u(x, y)$.

16. (*Three-dimensional case*) Consider the *three*-dimensional problem

$$\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0 \tag{16.1}$$

in the rectangular prism $0 < x < a$, $0 < y < b$, $0 < z < c$, where $u = 0$ on each of the six faces except for the face $z = c$, on which u is a prescribed function $f(x, y)$,

$$u(x, y, c) = f(x, y). \qquad (0 < x < a,\ 0 < y < b) \tag{16.2}$$

(a) Use separation of variables to derive the solution

$$u(x, y, z) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} D_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \sinh \omega_{mn} z, \tag{16.3}$$

where

$$\begin{aligned} \omega_{mn} &= \pi \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}, \\ D_{mn} &= \frac{4}{ab \sinh(\omega_{mn} c)} \int_0^b \int_0^a f(x, y) \\ &\qquad \times \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}\, dx\, dy. \end{aligned} \tag{16.4}$$

(b) For the case where $c = b = a$ and $f(x, y) = 100$, use (16.3) and (16.4) to evaluate $u(a/2, a/2, a/2)$.

17. (*A necessary condition for existence*) Consider the Poisson problem

$$\nabla^2 u = f(x, y, z) \tag{17.1}$$

in some three-dimensional domain $\mathcal{D}$ with surface $\mathcal{S}$. Integrating (17.1) over $\mathcal{D}$, show that

$$\int_{\mathcal{S}} \frac{\partial u}{\partial n}\, dA = \int_{\mathcal{D}} f\, dV. \tag{17.2}$$

NOTE: Thus, the boundary values of $\partial u/\partial n$ (whether they are specified or not) need to be consistent with the source f in the sense of (17.2) if a solution to (17.1) is to exist. For instance, suppose a homogeneous Neumann condition is appended to (17.1), that $\partial u/\partial n = 0$ everywhere on $\mathcal{S}$. Then, (17.2) tells us that for a solution to exist the net source must be zero: $\int_{\mathcal{D}} f\, dV = 0$. That result makes sense physically because if the integral were positive, say, then the average temperature within $\mathcal{D}$ would be an increasing function of time, whereas (17.1) is based on *steady-state* conduction.

18. (*Uniqueness*) Suppose that $u(x, y, z)$ is C^2 and satisfies the Poisson equation (17.1) throughout a domain $\mathcal{D}$, together with a Dirichlet boundary condition $u = g(x, y, z)$ on the (piecewise smooth orientable) surface $\mathcal{S}$ of $\mathcal{D}$.
(a) Show that that solution is unique. HINT: Suppose that there are two such solutions, say u_1 and u_2. With $w \equiv u_1 - u_2$, show that $\nabla^2 w = 0$ in $\mathcal{D}$ and $w = 0$ on $\mathcal{S}$. With "u" = "v" = w in Green's first identity, show that

$$\int_{\mathcal{D}} (w_x^2 + w_y^2 + w_z^2)\, dV = 0, \tag{18.1}$$

and conclude that $w_x = w_y = w_z$ so w is at most a constant. Show that the constant must be zero, so $u_1 = u_2$ in $\mathcal{D}$.
(b) Repeat (a) with the Dirichlet condition replaced by the Neumann condition $\partial u/\partial n = g$. This time, show that the solution is unique only to within an arbitrary additive constant.

(c) Repeat (a) with the Dirichlet condition replaced by a mixed boundary condition whereby $u = g$ over part of $\mathcal{S}$ and $\partial u/\partial n = h$ over the rest of $\mathcal{S}$.

19. (*Application of Sturm–Liouville theory*) (a) Solve the problem

$$\begin{gathered} u_{xx} + u_{yy} = 0, \\ u(0, y) = u(x, 0) = u(x, 3) + 5u_y(x, 3) = 0, \\ u(4, y) = 100 \end{gathered} \tag{19.1a,b,c}$$

on the rectangle $0 < x < 4$, $0 < y < 3$ by separation of variables, with the help of the Sturm–Liouville theory. Show that the values of the separation constant κ are the (nonzero) roots of the equation

$$\tan 3\kappa = -5\kappa, \tag{19.2}$$

and denote those roots as κ_n $(n = 1, 2, \ldots)$. Use computer software to evaluate κ_1 through κ_5 and use the first five terms of your series solution to estimate $u(2, 1)$.

(b) Same as (a), with (19.1c) changed to $u_x(4, y) = 100$.

20.3 Separation of Variables; Non-Cartesian Coordinates

20.3.1. Plane polar coordinates. Let r, θ be the usual plane polar coordinates with $x = r\cos\theta$ and $y = r\sin\theta$. If the problem domain is bounded by constant r and constant θ curves, then we must use r and θ as our independent variables because, for the separation of variable method to work, we need the boundary conditions to be given on the constant coordinate curves. Since we will need to express the Laplace equation in terms of r and θ, we recall from (24) in Section 16.7 that for plane polar coordinates the Laplacian is

$$\nabla^2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}. \tag{1}$$

EXAMPLE 1. Consider the Dirichlet problem

$$\nabla^2 u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0, \quad (a < r < b,\ 0 < \theta < \alpha) \tag{2a}$$

$$u(r, 0) = u_1, \qquad (a < r < b) \tag{2b}$$

$$u(r, \alpha) = u_2, \qquad (a < r < b) \tag{2c}$$

$$u(a, \theta) = 0, \qquad (0 < \theta < \alpha) \tag{2d}$$

$$u(b, \theta) = f(\theta) \qquad (0 < \theta < \alpha) \tag{2e}$$

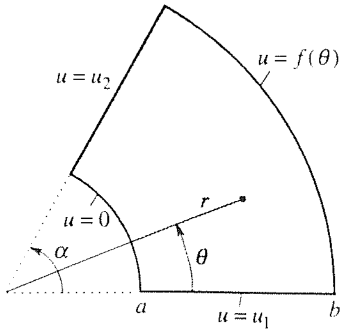


Figure 1. The Dirichlet problem (2).

shown in Fig. 1. Observe that this problem is similar to the basic Cartesian coordinate version (Fig. 1 of Section 20.2), with some "distortion," so we expect our solution steps to be similar as well.

According to the method of separation of variables, we seek u in the product form

$$u(r, \theta) = R(r)\Theta(\theta). \tag{3}$$

Putting (3) into (2a) gives

$$R''\Theta + \frac{1}{r}R'\Theta + \frac{1}{r^2}R\Theta'' = 0,$$

and if we multiply by r^2 and divide by $R\Theta$ to separate the variables we obtain

$$\frac{r^2R'' + rR'}{R} = -\frac{\Theta''}{\Theta} = \text{constant} = \kappa^2. \tag{4}$$

Here, we choose $+\kappa^2$ so as to obtain $\cos \kappa\theta$ and $\sin \kappa\theta$ solutions for Θ (rather than $\cosh \kappa\theta$ and $\sinh \kappa\theta$) since we anticipate that to satisfy the $u(b, \theta) = f(\theta)$ boundary condition we will need to expand $f(\theta)$ in a Fourier series.

Proceeding, the R and Θ equations are

$$r^2R'' + rR' - \kappa^2R = 0, \tag{5}$$

$$\Theta'' + \kappa^2\Theta = 0. \tag{6}$$

Although (5) has nonconstant coefficients (so the solution form $R = e^{\lambda r}$ will not work), it is elementary because it is of Cauchy–Euler form. Accordingly, seek $R = r^\lambda$ and obtain $\lambda(\lambda - 1) + \lambda - \kappa^2 = 0$ or $\lambda = \pm\kappa$. Thus, (5) admits two linearly independent solutions, $R = r^\kappa$ and $R = r^{-\kappa}$, unless $\kappa = 0$ in which case the two solutions coalesce into the one solution $R =$ constant. To find the missing solution, for the case $\kappa = 0$, put $\kappa = 0$ into (5) and obtain $r^2R'' + rR' = 0$. The latter can be reduced to the first-order equation $r\,dp/dr + p = 0$ by the substitution $R' = p$ and integrated to give $p(r) = C_1/r$. Thus, $R(r) = \int p\,dr = C_1 \ln r + C_2$, so we have these general solutions for R and Θ:

$$R(r) = \begin{cases} A + B\ln r, & \kappa = 0 \\ Cr^\kappa + Dr^{-\kappa}, & \kappa \neq 0 \end{cases} \tag{7}$$

$$\Theta(\theta) = \begin{cases} E + F\theta, & \kappa = 0 \\ G\cos\kappa\theta + H\sin\kappa\theta. & \kappa \neq 0 \end{cases} \tag{8}$$

Then, with the help of superposition, we have

$$u(r, \theta) = (A + B\ln r)(E + F\theta) + (Cr^\kappa + Dr^{-\kappa})(G\cos\kappa\theta + H\sin\kappa\theta). \tag{9}$$

Saving the boundary condition on $r = b$ for last, we first apply the boundary conditions on the adjacent edges $\theta = 0$ and $\theta = \alpha$:

$$u(r, 0) = u_1 = (A + B\ln r)E + (Cr^\kappa + Dr^{-\kappa})G, \tag{10}$$

so we set $AE = u_1$, $B = 0$, and $G = 0$. Updating (9) accordingly,

$$u(r, \theta) = u_1 + I\theta + (Pr^\kappa + Qr^{-\kappa})\sin\kappa\theta, \tag{11}$$

where we have combined AF as I, CH as P, and DH as Q. Next,

$$u(r, \alpha) = u_2 = u_1 + I\alpha + (Pr^\kappa + Qr^{-\kappa})\sin\kappa\alpha, \tag{12}$$

so $u_1 + I\alpha = u_2$ and $\sin\kappa\alpha = 0$, hence $I = (u_2 - u_1)/\alpha$ and $\kappa = n\pi/\alpha$ $(n = 1, 2, \dots)$. Thus,

$$u(r, \theta) = u_1 + (u_2 - u_1)\frac{\theta}{\alpha} + \sum_{n=1}^{\infty}\left(P_n r^{n\pi/\alpha} + Q_n r^{-n\pi/\alpha}\right)\sin\frac{n\pi\theta}{\alpha}. \tag{13}$$

Then

$$u(a,\theta) = 0 = u_1 + (u_2 - u_1)\frac{\theta}{\alpha} + \sum_{n=1}^{\infty}\left(P_n a^{n\pi/\alpha} + Q_n a^{-n\pi/\alpha}\right)\sin\frac{n\pi\theta}{\alpha}.$$

or, moving the known terms on the right to the left-hand side,

$$-u_1 - (u_2 - u_1)\frac{\theta}{\alpha} = \sum_{n=1}^{\infty}\left(P_n a^{n\pi/\alpha} + Q_n a^{-n\pi/\alpha}\right)\sin\frac{n\pi\theta}{\alpha}. \tag{14}$$

The latter is a half-range sine expansion of $-u_1 - (u_2 - u_1)(\theta/\alpha)$ so we can compute the coefficients $P_n a^{n\pi/\alpha} + Q_n a^{-n\pi/\alpha}$ as

$$P_n a^{n\pi/\alpha} + Q_n a^{-n\pi/\alpha} = \frac{2}{\alpha}\int_0^{\alpha}\left[-u_1 - (u_2 - u_1)\frac{\theta}{\alpha}\right]\sin\frac{n\pi\theta}{\alpha}\,d\theta. \tag{15}$$

Finally,

$$u(b,\theta) = f(\theta) = u_1 + (u_2 - u_1)\frac{\theta}{\alpha} + \sum_{n=1}^{\infty}\left(P_n b^{n\pi/\alpha} + Q_n b^{-n\pi/\alpha}\right)\sin\frac{n\pi\theta}{\alpha},$$

or

$$f(\theta) - u_1 - (u_2 - u_1)\frac{\theta}{\alpha} = \sum_{n=1}^{\infty}\left(P_n b^{n\pi/\alpha} + Q_n b^{-n\pi/\alpha}\right)\sin\frac{n\pi\theta}{\alpha}, \tag{16}$$

so

$$P_n b^{n\pi/\alpha} + Q_n b^{-n\pi/\alpha} = \frac{2}{\alpha}\int_0^{\alpha}\left[f(\theta) - u_1 - (u_2 - u_1)\frac{\theta}{\alpha}\right]\sin\frac{n\pi\theta}{\alpha}\,d\theta. \tag{17}$$

We can evaluate the integrals in (15) and (17), once $f(\theta)$ is specified, so (15) and (17) amount to two linear algebraic equations in the unknown P_n's and Q_n's. They have a unique solution because the determinant of the coefficient matrix is

$$\begin{vmatrix} a^{n\pi/\alpha} & a^{-n\pi/\alpha} \\ b^{n\pi/\alpha} & b^{-n\pi/\alpha} \end{vmatrix} = \left(\frac{a}{b}\right)^{n\pi/\alpha} - \left(\frac{b}{a}\right)^{n\pi/\alpha} \neq 0 \tag{18}$$

since $b \neq a$. Thus, the solution to (2) is given by (13), with P_n and Q_n determined from (15) and (17). For instance, if $u_1 = u_2 = 0$ and $f(\theta) = 100$, then we obtain (Exercise 1)

$$u(r,\theta) = \frac{400}{\pi}\sum_{1,3,\dots}^{\infty}\frac{1}{n}\,\frac{\left(\dfrac{r}{a}\right)^{n\pi/\alpha} - \left(\dfrac{a}{r}\right)^{n\pi/\alpha}}{\left(\dfrac{b}{a}\right)^{n\pi/\alpha} - \left(\dfrac{a}{b}\right)^{n\pi/\alpha}}\sin\frac{n\pi\theta}{\alpha}. \tag{19}$$

COMMENT 1. The Laplace equation in polar coordinates, (2a), did indeed prove to be separable. That is, putting $u(r,\theta) = R(r)\Theta(\theta)$ into (2a) we were able to get all of the r dependence on one side of the equation and all of the θ dependence on the other [in (4)], and hence to infer the ODE's (5) and (6) on R and Θ. In fact, the diffusion, wave, and Laplace equations are all separable in Cartesian, polar, cylindrical, and spherical coordinates, for which we can be grateful because *not all PDE's are separable*; see, for instance, Exercise 14 in Section 20.2.

COMMENT 2. Given that the Fourier expansion will take place on the $r = b$ edge, there is no extra difficulty in admitting $u =$ constant boundary conditions on the adjacent edges, namely, the conditions $u = u_1$ on $\theta = 0$ and $u = u_2$ on $\theta = \alpha$. The reason is that the $A(E + F\theta)$ part of the solution (9), which comes from $\kappa = 0$, is able to handle those boundary conditions; it gives the $u_1 + (u_2 - u_1)\theta/\alpha$ part of the final solution (13). If we spoke of the analogous $(A + Bx)(E + Fy)$ term in Section 20.2 as being a "ramp" function (if B or F is zero), we might call the $A(E + F\theta)$ term a "fan" function since its graph "fans" from the value u_1 at one value of θ to u_2 at another value of θ (Fig. 2).

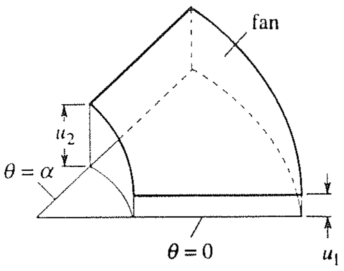


Figure 2. Fan from u_1 to u_2.

COMMENT 3. If the boundary condition along either radial edge ($\theta = 0$ or $\theta = \alpha$) is nonconstant, so that a Fourier expansion is needed along that edge, then the solution is more difficult because the needed expansion turns out not to be a familiar half- or quarter-range Fourier series but rather a Sturm–Liouville eigenfunction expansion. The solution for that case is outlined in Exercise 13.

COMMENT 4. The boundary conditions (2b)–(2e) are of Dirichlet type. What if they were of Neumann type ($\partial u/\partial n$ prescribed)? For example, if (2b) were replaced by the Neumann condition

$$\frac{\partial u}{\partial n}(r, 0) = g(r), \qquad (a < r < b) \tag{20}$$

where $g(r)$ is prescribed, then what is $\partial u/\partial n$ in terms of r and θ? The key is to use the directional derivative formula $du/ds = \nabla u \cdot \hat{\mathbf{s}}$ in Section 16.4, which gives

$$\frac{\partial u}{\partial n} = \nabla u \cdot \hat{\mathbf{n}} = \left(\frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta \right) \cdot (-\hat{\mathbf{e}}_\theta) = -\frac{1}{r}\frac{\partial u}{\partial \theta}$$

on the $\theta = 0$ edge, so (20) can be expressed in terms of r, θ as $-(1/r)\partial u/\partial\theta = g(r)$ or

$$\frac{\partial u}{\partial \theta}(r, 0) = -rg(r). \qquad (a < r < b) \tag{21}$$

Similarly, a Neumann condition

$$\frac{\partial u}{\partial n}(a, \theta) = h(\theta) \qquad (0 < \theta < \alpha) \tag{22}$$

becomes

$$\frac{\partial u}{\partial r}(a, \theta) = -h(\theta) \qquad (0 < \theta < \alpha) \tag{23}$$

because

$$\frac{\partial u}{\partial n} = \nabla u \cdot \hat{\mathbf{n}} = \left(\frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta \right) \cdot (-\hat{\mathbf{e}}_r) = -\frac{\partial u}{\partial r}$$

on $r = a$. Physically, remember that if u is a temperature field then $\partial u/\partial n$ is proportional to the *heat flux across* that boundary. ▮

EXAMPLE 2. *Dirichlet Problem for Circular Disk.* Next, consider the Dirichlet problem

$$\nabla^2 u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0, \qquad (0 \le r < b) \tag{24a}$$

$$u(b,\theta) = f(\theta) \qquad (-\infty < \theta < \infty) \tag{24b}$$

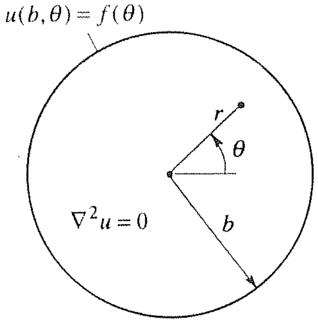


Figure 3. Dirichlet problem for disk.

shown in Fig. 3. We begin with (9) and write

$$u(r,\theta) = (A + B\ln r)(E + F\theta) + (Cr^{\kappa} + Dr^{-\kappa})(G\cos\kappa\theta + H\sin\kappa\theta), \tag{25}$$

but comparing (24) with (2) we seem to be missing some boundary conditions. After all, the PDE is of second order in r and in θ, so we expect to need two r boundary conditions and two θ boundary conditions, as were present in (2b)–(2e). For insight, it is useful to imagine obtaining the disk in Fig. 3 as the limiting case of the region in Fig. 1 as $a \to 0$ and $\alpha \to 2\pi$. Letting $a \to 0$ first, observe that the $r = a$ boundary curve shrinks to a point; when that happens we lose that boundary curve and corresponding boundary condition and have a pie-shaped region. Next, let $\alpha \to 2\pi$. The moment α becomes 2π we lose the two boundary conditions on the edges $\theta = 0$ and $\theta = \alpha$ because those edges disappear as boundary edges and become interior to the region.

To compensate for these losses we do the following. First, we adjoin to (24a,b), in lieu of the missing r boundary condition, a *boundedness condition* at $r = 0$, namely,

$$u(r,\theta) \quad \text{bounded as} \quad r \to 0. \tag{26}$$

To apply this condition, observe that both the $\ln r$ and $r^{-\kappa}$ terms in (25) are unbounded as $r \to 0$: $\ln r \to -\infty$ and $r^{-\kappa} \to \infty$ as $r \to 0$. Thus, (26) requires us to remove those terms by setting $B = D = 0$, in which case (25) reduces to

$$u(r,\theta) = I + J\theta + r^{\kappa}(P\cos\kappa\theta + Q\sin\kappa\theta). \tag{27}$$

To remedy the situation regarding the missing θ boundary condition we begin by observing a key difference between Example 1 and the present example: the θ domain in Example 1 was finite $(0 < \theta < \alpha)$ whereas here it is infinite $(-\infty < \theta < \infty)$, for we see from Fig. 3 that there is nothing to prevent the representative point from encircling the origin repeatedly, clockwise or counterclockwise. Thus, if $u(r,\theta)$ is to be a single-valued function of θ and hence uniquely defined at each point within the disk, then it needs to be 2π-periodic in θ:

$$u(r,\theta + 2\pi) = u(r,\theta). \tag{28}$$

This periodicity will compensate for the two missing θ boundary conditions (see also Exercise 8) so the full problem is given by (24a), (24b), (26), and (28).

Let us impose (28) on each term in (27). First, I is a constant and is therefore 2π-periodic; hence, retain that term. Next, $J\theta$ is *not* periodic (as can be seen from its linear graph), so we must set $J = 0$ to remove that term. Finally, the $\cos\kappa\theta$ and $\sin\kappa\theta$ terms are periodic, but we need to determine the allowable κ's so that they are **2π-periodic.** According to the definition of periodicity, we need

$$\cos\kappa(\theta + 2\pi) = \cos\kappa\theta \tag{29}$$

for all θ (and similarly for the sine term) or, since $\cos(A + B) = \cos A\cos B - \sin A\sin B$, we need

$$\cos\kappa\theta\cos 2\pi\kappa - \sin\kappa\theta\sin 2\pi\kappa = \cos\kappa\theta.$$

Equating coefficients of the linearly independent $\cos \kappa\theta$ and $\sin \kappa\theta$ terms gives $\cos 2\pi\kappa = 1$ and $\sin 2\pi\kappa = 0$, with the roots*

$$\cos 2\pi\kappa = 1: \quad \kappa = 1, 2, 3, \ldots, \tag{30a}$$

$$\sin 2\pi\kappa = 0: \quad \kappa = \frac{1}{2}, 1, \frac{3}{2}, 2, \ldots. \tag{30b}$$

Since both conditions need to hold, we accept only κ's that are in both lists, namely, $\kappa = 1, 2, 3, \ldots$. The same result is obtained when we enforce the 2π-periodicity of the $\sin \kappa\theta$ term.

Thus, with $J = 0$, $\kappa = n$, and with the help of superposition, (26) gives

$$u(r, \theta) = I + \sum_{n=1}^{\infty} r^n (P_n \cos n\theta + Q_n \sin n\theta), \tag{31}$$

We are ready for the final boundary condition:

$$u(b, \theta) = f(\theta) = I + \sum_{n=1}^{\infty} b^n (P_n \cos n\theta + Q_n \sin n\theta), \tag{32}$$

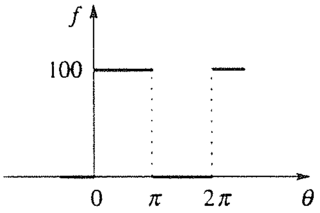


Figure 4. Square wave f.

which holds on $-\infty < \theta < \infty$. Notice carefully that whereas (14) and (16) are half-range sine expansions on the finite interval $0 < \theta < \alpha$, (32) is the full Fourier series expansion of the 2π-periodic function $f(\theta)$ on $-\infty < \theta < \infty$. Accordingly [see (5) in Section 17.3, with $\ell = \pi$],

$$I = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta)\, d\theta, \qquad P_n = \frac{1}{\pi b^n} \int_{-\pi}^{\pi} f(\theta) \cos n\theta \, d\theta,$$
$$Q_n = \frac{1}{\pi b^n} \int_{-\pi}^{\pi} f(\theta) \sin n\theta \, d\theta, \tag{33}$$

and the solution is given by (31) and (33).

To illustrate, let the boundary temperature $f(\theta)$ be 100 on the upper half of the circle and 0 on the lower half, in which case f is actually the 2π-periodic square wave shown in Fig. 4. Using (33), the result is

$$u(r, \theta) = 50 + \frac{200}{\pi} \sum_{1,3,\ldots}^{\infty} \left(\frac{r}{b}\right)^n \frac{\sin n\theta}{n}, \tag{34}$$

and representative isotherms are shown in Fig. 5.

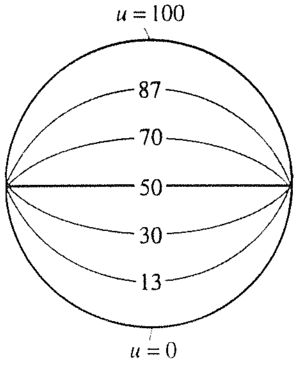


Figure 5. Isotherms corresponding to (34).

COMMENT 1. Setting $r = 0$ in (31), observe that u at the center of the disk equals I, and I, according to (33), is the average value of the boundary temperature . For the example shown in Fig. 5, for instance, $f(\theta)$ is 100 on the upper half of the circumference and 0 on

*As usual, we exclude negative κ values since they contribute nothing new. For instance, if we change κ to $-\kappa$ in the last term in (25) then that term takes the equivalent form $(Cr^{-\kappa} + Dr^{\kappa})(G \cos \kappa\theta - H \sin \kappa\theta)$. This result is not surprising since the κ^2 in (4) cannot distinguish between positive κ's and negative κ's. Further, we disallow $\kappa = 0$ in (30) because the $\kappa = 0$ case is handled separately, in (25), by the $(A + B \ln r)(E + F\theta)$ term.

the lower half so the average value is 50. Sure enough, the isotherm $u = 50$ does pass through the origin in Fig. 5.

COMMENT 2. Let us put (33) into (31). First changing the dummy variable of integration to ϑ, say, to avoid confusion with the θ's in (31), the result is

$$u(r,\theta) = \frac{1}{\pi}\int_{-\pi}^{\pi}\left[\frac{1}{2} + \sum_{n=1}^{\infty}\left(\frac{r}{b}\right)^n \cos n(\vartheta - \theta)\right] f(\vartheta)\, d\vartheta, \tag{35}$$

where we have formally interchanged the order of integration and summation. It is striking that the infinite series in (35) can be summed, that is, gotten into closed form. The result (Exercise 9) is

$$\boxed{\begin{aligned} u(r,\theta) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{b^2 - r^2}{b^2 - 2br\cos(\vartheta - \theta) + r^2} f(\vartheta)\, d\vartheta \\ &\equiv \textstyle\int_{-\pi}^{\pi} P(r, \vartheta - \theta) f(\vartheta)\, d\vartheta; \end{aligned}} \tag{36}$$

(36) is the **Poisson integral formula for the circular disk** and $P(r, \vartheta - \theta)$ is the corresponding Poisson kernel. ▮

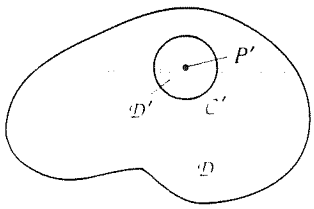


Figure 6. Average value theorem.

In Comment 1, above, we wrote that the temperature u at the center of the disk equals the average of the boundary temperatures $f(\theta)$ around the circumference. This result can be generalized as follows. Within an arbitrary domain $\mathcal{D}$, not necessarily circular (Fig. 6), consider any point P' and any circular domain $\mathcal{D}'$ that is centered at P' and that lies entirely within $\mathcal{D}$, and suppose that $\nabla^2 u = 0$ in $\mathcal{D}$. Then, whatever the temperatures are on the boundary C' of $\mathcal{D}'$ we can consider them as boundary conditions for the sub-problem $\nabla^2 u = 0$ in $\mathcal{D}'$, which problem was the subject of Example 2. From the average value result found in Example 2 we know that u at P' is the average of the u values around C'. Thus, *if $\nabla^2 u = 0$ in a two-dimensional domain $\mathcal{D}$, then the temperature u at any point P' within $\mathcal{D}$ is equal to the average temperature around any circle centered at P' and lying within $\mathcal{D}$.* This result, known as the **average value property** of the Laplace equation, is said to be a "local" result since it holds for an arbitrarily small circle C' and is insensitive to the shape of $\mathcal{D}$. It holds in one dimension (Exercise 10) and in three dimensions as well (as we will see in Example 5).

The average value property enables us to prove the **maximum principle** for the Laplace equation, which is as follows: *Let u be the steady-state temperature field within a two-dimensional domain $\mathcal{D}$, so u satisfies the Laplace equation $\nabla^2 u = 0$. Then u cannot attain its maximum value in $\mathcal{D}$ (unless u is a constant everywhere); it must attain its maximum on the boundary of $\mathcal{D}$.* For suppose that u does have a maximum value M, say, at a point P within $\mathcal{D}$. Since u at P equals the average value of u around any circle centered at P, u must achieve values less than M and greater than M at points within $\mathcal{D}$ if it is not simply a constant everywhere. But that result contradicts our assumption that the maximum value of u is M, hence u cannot have a maximum value within $\mathcal{D}$. By virtually the same argument we can show that u cannot attain its minimum value in $\mathcal{D}$, and we obtain the **minimum principle**, which is identical to the maximum principle (italicized above) with the two "maximums" changed to "minimums." It follows from these maximum and minimum principles that the values of u within $\mathcal{D}$ necessarily lie between the minimum and maximum values of u on the boundary of $\mathcal{D}$.

These results make sense physically, for suppose that the temperature maintained on the boundary of $\mathcal{D}$ lies between 50°C and 70°C and let $u = 2{,}000$°C at some point inside $\mathcal{D}$. Surely that "hot spot" will cool down, with time, and the surrounding material will heat up. But that is impossible for we have assumed steady-state heat conduction; that is, the Laplace equation $\nabla^2 u = 0$ is the steady-state version of the heat conduction equation $\alpha^2\nabla^2 u = u_t$.

20.3.2. Cylindrical coordinates. (Optional) The steady-state temperature field within a cylindrical rod (Fig. 7) is governed by the Laplace equation in cylindrical coordinates,

$$\nabla^2 u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} + u_{zz} = 0. \tag{37}$$

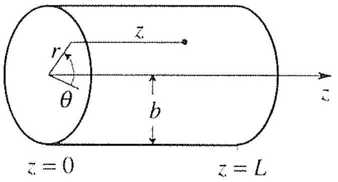


Figure 7. Cylindrical rod.

Suppose there is axisymmetry so that u does not vary with θ. Then (37) reduces to

$$\nabla^2 u = u_{rr} + \frac{1}{r}u_r + u_{zz} = 0, \tag{38}$$

which case we consider here. Specifically, we consider the problem shown in the left-hand member of Fig. 8, and we begin by breaking it into the two problems shown in the figure. We will solve the u_1 problem as Example 3 and the u_2 problem as Example 4.

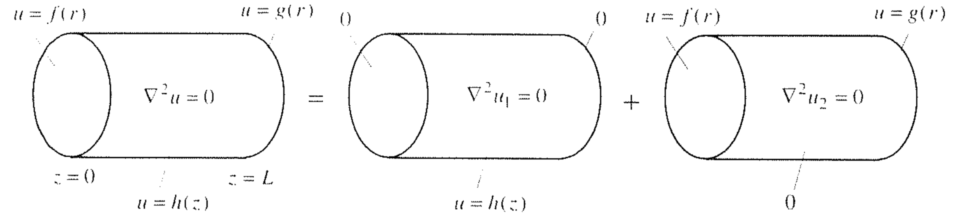


Figure 8. Breakdown by superposition.

EXAMPLE 3. *The u_1 Problem.* To solve (38), seek

$$u_1(r, z) = R(r)Z(z). \tag{39}$$

Putting the latter into (38) and separating variables gives

$$\frac{R'' + \frac{1}{r}R'}{R} = -\frac{Z''}{Z} = \text{constant} = \kappa^2 \tag{40}$$

and the ODE's

$$R'' + \frac{1}{r}R' - \kappa^2 R = 0, \tag{41}$$

$$Z'' + \kappa^2 Z = 0. \tag{42}$$

We chose the $+\kappa^2$ in (40) so as to obtain $\cos\kappa z$ and $\sin\kappa z$ (rather than $\cosh\kappa z$ and $\sinh\kappa z$) solutions of the Z equation since we look ahead to a Fourier series expansion of $h(z)$. Distinguishing the cases $\kappa = 0$ and $\kappa \neq 0$, the general solutions of (41) and (42) are (Exercise 14):

$$R = \begin{cases} A + B\ln r, & \kappa = 0 \\ CI_0(\kappa r) + DK_0(\kappa r), & \kappa \neq 0 \end{cases} \tag{43}$$

$$Z = \begin{cases} E + Fz, & \kappa = 0 \\ G\cos\kappa z + H\sin\kappa z, & \kappa \neq 0 \end{cases} \tag{44}$$

where I_0, K_0 are modified Bessel functions of the first and second kind, respectively, of order zero. Thus,

$$u_1(r,z) = (A + B\ln r)(E + Fz) + [CI_0(\kappa r) + DK_0(\kappa r)](G\cos\kappa z + H\sin\kappa z). \tag{45}$$

In the z variable we have the two boundary conditions at $z = 0$ and $z = L$, but in r we have only the boundary condition at $r = b$ so in lieu of a second r boundary condition we require that u be bounded as $r \to 0$ (i.e., all along the z axis). Since $\ln r \to -\infty$ we set $B = 0$, and since $K_0(\kappa r) \sim -\ln r \to \infty$ as $r \to 0$ (Fig. 9) we set $D = 0$, so (45) becomes

$$u_1(r,z) = P + Qz + I_0(\kappa r)(S\cos\kappa z + T\sin\kappa z), \tag{46}$$

where we have combined AE as P, AF as Q, CG as S, and CH as T, for brevity.

Next, we apply the conditions at $z = 0$, $z = L$, and $r = b$, in turn:

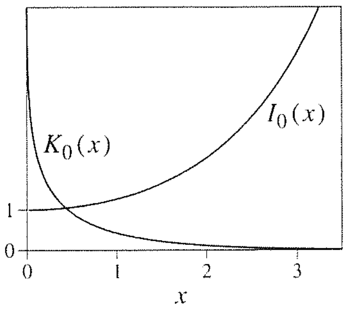


Figure 9. I_0 and K_0.

$$u_1(r,0) = 0 = P + I_0(\kappa r)S, \tag{47}$$

so $P = 0$ and $S = 0$. Updating (46) accordingly,

$$u_1(r,z) = Qz + TI_0(\kappa r)\sin\kappa z. \tag{48}$$

Next,

$$u_1(r,L) = 0 = QL + TI_0(\kappa r)\sin\kappa L, \tag{49}$$

so $Q = 0$ and $\kappa = n\pi/L$ for $n = 1, 2, \dots$. Updating (48),

$$u_1(r,z) = \sum_{n=1}^{\infty} T_n I_0\left(\frac{n\pi r}{L}\right)\sin\frac{n\pi z}{L}. \tag{50}$$

Finally,

$$u_1(b,z) = h(z) = \sum_{n=1}^{\infty} T_n I_0\left(\frac{n\pi b}{L}\right)\sin\frac{n\pi z}{L}, \qquad (0 < z < L) \tag{51}$$

which is a half-range sine series so

$$T_n I_0\left(\frac{n\pi b}{L}\right) = \frac{2}{L}\int_0^L h(z)\sin\frac{n\pi z}{L}\,dz,$$

or

$$T_n = \frac{2}{LI_0(n\pi b/L)}\int_0^L h(z)\sin\frac{n\pi z}{L}\,dz. \tag{52}$$

The desired solution is given by (51) and (52).

COMMENT. As in Example 2 we applied the boundedness condition first. Indeed, whenever there is a boundedness condition we suggest that you apply it first because it will eliminate one or more terms, thereby giving an immediate simplification of the solution form. ▮

EXAMPLE 4. *The u_2 Problem.* To solve for u_2 (see Fig. 8) we again seek

$$u_2(r, z) = R(r)Z(z) \tag{53}$$

and obtain

$$\frac{R'' + \frac{1}{r}R'}{R} = -\frac{Z''}{Z} = \text{constant} = -\kappa^2 \tag{54}$$

and the ODE's

$$R'' + \frac{1}{r}R' + \kappa^2 R = 0, \tag{55}$$

$$Z'' - \kappa^2 Z = 0, \tag{56}$$

where this time we chose $-\kappa^2$, so as to obtain oscillatory solutions of the R equation (rather than the I_0, K_0 pair obtained in Example 3). Specifically (Exercise 15),

$$R = \begin{cases} A + B\ln r, & \kappa = 0 \\ CJ_0(\kappa r) + DY_0(\kappa r), & \kappa \neq 0 \end{cases} \tag{57}$$

$$Z = \begin{cases} E + Fz, & \kappa = 0 \\ G\cosh\kappa z + H\sinh\kappa z, & \kappa \neq 0 \end{cases} \tag{58}$$

where J_0, Y_0 are Bessel functions of the first and second kind, respectively, of order zero. Thus,

$$u_2(r, z) = (A + B\ln r)(E + Fz) + [CJ_0(\kappa r) + DY_0(\kappa r)](G\cosh\kappa z + H\sinh\kappa z). \tag{59}$$

Boundedness as $r \to 0$ requires that $B = 0$ and $D = 0$, since $Y_0(\kappa r) \sim (2/\pi)\ln r \to -\infty$ as $r \to 0$ (Fig. 10), so (59) reduces to

$$u_2(r, z) = P + Qz + J_0(\kappa r)(S\cosh\kappa z + T\sinh\kappa z). \tag{60}$$

Since we look ahead to expanding $f(r)$ and $g(r)$, we must apply both r boundary conditions first, before attempting either of the end conditions at $z = 0$ and L. Having already applied the boundedness condition at $r = 0$, we next write

$$u_2(b, z) = 0 = P + Qz + J_0(\kappa b)(S\cosh\kappa z + T\sinh\kappa z),$$

which requires that $P = 0$, $Q = 0$, and

$$J_0(\kappa b) = 0 \tag{61}$$

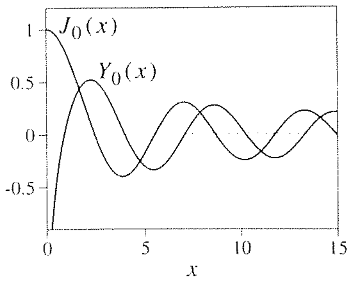


Figure 10. J_0 and Y_0.

[since we cannot afford to set $S = T = 0$ and lose the entire $J_0(\kappa r)(S \cosh \kappa z + T \sinh \kappa z)$ term in (60)]. Denoting the roots of $J_0(x) = 0$ as $z_1, z_2, \ldots$ (z's because they give the *zeros* of J_0), we learn from (61) that $\kappa = z_n/b$ $(n = 1, 2, \ldots)$. Updating (60) accordingly,

$$u_2(r, z) = \sum_{n=1}^{\infty} J_0\left(z_n \frac{r}{b}\right)\left[S_n \cosh\left(z_n \frac{z}{b}\right) + T_n \sinh\left(z_n \frac{z}{b}\right)\right]. \tag{62}$$

Finally, the end conditions give

$$u_2(r, 0) = f(r) = \sum_{n=1}^{\infty} S_n J_0\left(z_n \frac{r}{b}\right), \tag{63}$$

and

$$u_2(r, L) = g(r) = \sum_{n=1}^{\infty}\left[S_n \cosh\left(z_n \frac{L}{b}\right) + T_n \sinh\left(z_n \frac{L}{b}\right)\right] J_0\left(z_n \frac{r}{b}\right) \tag{64}$$

on $0 < r < b$. How are we to solve (63) and (64) for S_n and T_n? There are two questions raised by (63): first, is it *possible* to expand a given function $f(r)$ on the interval $0 < r < b$ in the form of an infinite linear combination of $J_0(z_n r/b)$ terms and, second, if so, how do we compute the S_n coefficients? Similarly for (64). Both questions are answered by the Sturm–Liouville theory, for the problem

$$(rR')' + \kappa^2 rR = 0, \qquad (0 < r < b) \tag{65a}$$

$$R(0) \text{ bounded}, \quad R(b) = 0 \tag{65b}$$

governing R is a Sturm–Liouville problem, where "λ" $= \kappa^2$. This problem is studied in Example 2 of Section 17.8 so, referring you to that example for the details, we can conclude from (63) that

$$S_n = \frac{2}{b^2[J_1(z_n)]^2}\int_0^b f(r) J_0\left(z_n \frac{r}{b}\right) r\, dr, \tag{66}$$

and from (64) that

$$S_n \cosh\left(z_n \frac{L}{b}\right) + T_n \sinh\left(z_n \frac{L}{b}\right) = \frac{2}{b^2[J_1(z_n)]^2}\int_0^b g(r) J_0\left(z_n \frac{r}{b}\right) r\, dr. \tag{67}$$

Thus, the solution is obtained by solving (66) and (67) for S_n and T_n (once f and g are specified), and putting these values into (62).

COMMENT. Observe how the problem is "self-contained": how to carry out the necessary expansions (63) and (64) is fully explained by the Sturm–Liouville problem on R that is "built right in." Likewise in Example 3, although we did not mention it because we merely noticed that (51) is a half-range sine expansion. Alternatively, we could have used the Sturm–Liouville theory there too. Specifically, the relevant Sturm–Liouville problem there is

$$Z'' + \kappa^2 Z = 0, \qquad (0 < z < L) \tag{68a}$$

$$Z(0) = 0, \quad Z(L) = 0, \tag{68b}$$

with the eigenfunctions $\sin(n\pi z/L)$, so (51) gives

$$T_n I_0\left(\frac{n\pi b}{L}\right) = \frac{\langle h(z), \sin\frac{n\pi z}{L}\rangle}{\langle \sin\frac{n\pi z}{L}, \sin\frac{n\pi z}{L}\rangle}$$
$$= \frac{\int_0^L h(z)\sin\frac{n\pi z}{L}\,dz}{\int_0^L \sin^2\frac{n\pi z}{L}\,dz} = \frac{2}{L}\int_0^L h(z)\sin\frac{n\pi z}{L}\,dz,$$

which result agrees with (52). Note that the inner product weight function is r in (65) and 1 in (68).

Likewise in Example 1. We deduced (15) from (14) and (17) from (16) by noticing that (14) and (16) were half-range sine expansions of their left-hand sides, over $0 < \theta < \alpha$. Alternatively, we could have used the Sturm–Liouville theory. Since the expansions were in the θ variable, look to the Θ problem for the Sturm–Liouville problem, namely,

$$\Theta'' + \kappa^2\Theta = 0, \qquad (0 < \theta < \alpha) \tag{69a}$$
$$\Theta(0) = 0, \quad \Theta(\alpha) = 0, \tag{69b}$$

with eigenfunctions $\sin(n\pi\theta/\alpha)$ and weight function 1. Where do we get the homogeneous boundary conditions (69b) from, considering that $u(r, 0) = u_1$ and $u(r, \alpha) = u_2$ are, in general, nonzero? If we retrace the solution steps, beginning with (9), we find that the $G\cos\kappa\theta + H\sin\kappa\theta$ factor in (9), which contributes the $\sin(n\pi\theta/\alpha)$ eigenfunctions, satisfies the homogeneous boundary conditions (69b), with the $(A + B\ln r)(E + F\theta)$ term handling the u_1 and u_2 values. And, of course, we can see directly that $\sin(n\pi\theta/\alpha)$ does indeed vanish at $\theta = 0$ and at $\theta = \alpha$. ▮

Then the solution to the problem on u that is shown in the left-hand part of Fig. 8 is $u(r, z) = u_1(r, z) + u_2(r, z)$.

20.3.3. Spherical coordinates. (Optional) If the domain under consideration is bounded by constant ρ, ϕ, θ surfaces, then we need to work with the Laplace equation in spherical coordinates,

$$\nabla^2 u = \frac{1}{\rho^2}\left[\frac{\partial}{\partial\rho}\left(\rho^2\frac{\partial u}{\partial\rho}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial\phi}\left(\sin\phi\frac{\partial u}{\partial\phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial\theta^2}\right] = 0. \tag{70}$$

Let us restrict our attention to cases where u is axisymmetric about the polar axis z, that is, where u does not vary with θ. Then the $u_{\theta\theta}$ term in (70) is zero and (70) reduces to the PDE

$$\nabla^2 u = u_{\rho\rho} + \frac{2}{\rho}u_\rho + \frac{1}{\rho^2}u_{\phi\phi} + \frac{\cot\phi}{\rho^2}u_\phi = 0 \tag{71}$$

on $u(\rho, \phi)$.

EXAMPLE 5. *Dirichlet Problem for Sphere.* Consider the Dirichlet problem consisting of the PDE (71) in the sphere $0 \leq \rho < c$, with the boundary condition

$$u(c, \phi) = f(\phi) \qquad (0 \leq \phi \leq \pi) \tag{72}$$

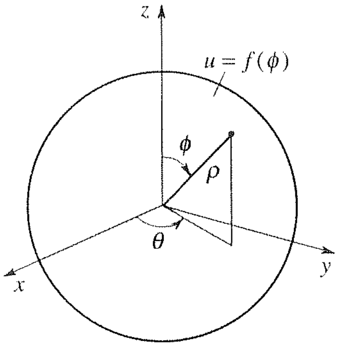


Figure 11. Dirichlet problem for sphere.

and the stipulation that u be bounded in the given domain (Fig. 11). This problem is the three-dimensional analog of the Dirichlet problem for a circular disk, which was the subject of Example 2.

To solve, seek

$$u(\rho, \phi) = R(\rho)\Phi(\phi) \tag{73}$$

and obtain, from (71),

$$\frac{\rho^2 R'' + 2\rho R'}{R} = -\frac{\Phi'' + \cot\phi\Phi'}{\Phi} = \text{constant} = \kappa^2, \tag{74}$$

and the ODE's

$$\rho^2 R'' + 2\rho R' - \kappa^2 R = 0, \tag{75}$$

$$\Phi'' + \cot\phi\,\Phi' + \kappa^2\Phi = 0. \tag{76}$$

The change of variables

$$\mu = \cos\phi$$

in (76) reduces that equation (Exercise 17) to the Legendre equation

$$(1-\mu^2)\frac{d^2\Phi}{d\mu^2} - 2\mu\frac{d\Phi}{d\mu} + \kappa^2\Phi = 0. \tag{77}$$

From our study of the Legendre equation in Section 4.4, we know that to obtain solutions of (77) that are bounded on $-1 \leq \mu \leq 1$ we need

$$\kappa^2 = n(n+1), \qquad (n = 0, 1, 2, \ldots) \tag{78}$$

in which case the corresponding bounded solutions are the Legendre polynomials

$$\Phi = P_n(\mu) = P_n(\cos\phi). \tag{79}$$

In terms of the physical domain, $\mu = 1$ corresponds to $\phi = 0$ so that unboundedness of Φ at $\mu = 1$ would mean unboundedness of the solution u all along the z axis from the "north pole" to the origin. Similarly, unboundedness of Φ at $\mu = -1$ would mean unboundedness of u along the z axis from the origin to the "south pole."

Turning to the R equation, with $\kappa^2 = n(n+1)$, the general solution of (75), which is a Cauchy–Euler equation, is

$$R(\rho) = A\rho^n + \frac{B}{\rho^{n+1}}. \qquad (n = 0, 1, 2, \ldots) \tag{80}$$

Here, the stipulated boundedness of u requires that $B = 0$. Putting (79) and (80) together and using superposition we have, thus far,

$$u(\rho, \phi) = \sum_{n=0}^{\infty} A_n \rho^n P_n(\cos\phi). \tag{81}$$

Finally, the boundary condition (72) requires that

$$u(c,\phi) = f(\phi) = \sum_{n=0}^{\infty} A_n c^n P_n(\cos\phi). \qquad (0 \le \phi \le \pi) \tag{82}$$

Since (82) involves an expansion in the ϕ variable, let us examine the boundary-value problem on Φ, namely,*

$$((1-\mu^2)\Phi')' + \kappa^2\Phi = 0, \qquad (-1 \le \mu \le 1) \tag{83a}$$

$$\Phi(-1) \text{ and } \Phi(1) \text{ finite.} \tag{83b}$$

Thus, we see that (82) amounts to a Fourier–Legendre expansion of the given function f in terms of the orthogonal eigenfunctions $P_n(\mu)$ or $P_n(\cos\phi)$, as was illustrated in Example 3 of Section 17.8. Accordingly,

$$A_n c^n = \frac{\langle f, P_n(\mu)\rangle}{\langle P_n(\mu), P_n(\mu)\rangle} = \frac{\int_{-1}^{1} f P_n\, d\mu}{\int_{-1}^{1} P_n^2\, d\mu} = \frac{2n+1}{2}\int_{-1}^{1} f P_n\, d\mu, \tag{84}$$

or

$$A_n = \frac{2n+1}{2c^n}\int_0^{\pi} f(\phi) P_n(\cos\phi)\sin\phi\, d\phi. \tag{85}$$

Hence, the solution is given by (81) and (85).

COMMENT. Observe that the value of u at the center of the sphere is

$$\begin{aligned} u(0,\phi) &= A_0 \qquad \text{[from (81)]} \\ &= \frac{1}{2}\int_{-1}^{1} f\, d\mu, \qquad \text{[from (84)]} \end{aligned}$$

which is the average value of the boundary temperature f. Thus, the average value property of the Laplace equation, discussed above, holds in three dimensions as well as two, and similarly for the maximum principle. ▮

Closure. We see, in Section 20.3.1, that the Laplace equation in plane polar coordinates is successfully separated and that although the ODE on $R(r)$ has nonconstant coefficients it is nevertheless an elementary equation, a Cauchy–Euler equation. We derive, as a result of Example 2, the average value property of the Laplace equation in two dimensions and the maximum principle as well. In Sections 20.3.2 and 20.3.3 we find that the Laplace equation can be separated in cylindrical and spherical coordinates as well, but that not all of the resulting ODE's are elementary: in cylindrical coordinates the $R(r)$ equation gives Bessel functions, and in spherical coordinates the $\Phi(\phi)$ equation gives Legendre polynomials.

*Strictly speaking, we should use a new name, such as $\Phi(\phi) = \Phi(\phi(\mu)) \equiv \Psi(\mu)$, say, but for economy of notation we use Φ whether the independent variable is ϕ or μ.

Expansions in Examples 4 and 5 involve the Bessel and Legendre functions, and they are carried out by relying on the Sturm–Liouville theory. If the expansion is on the r variable, for example, then we bring to light the Sturm–Liouville problem on $R(r)$ and use that problem and the Sturm–Liouville theory to guide our expansion.

Until now, our rule of thumb for choosing the sign of the κ^2 separation constant has been to choose the sign that makes oscillatory functions available for the eventual Fourier series expansion. If you are familiar with the Sturm–Liouville theory it would be helpful to make that rule more explicit as follows. Choose the sign of κ^2 so that the sign of the last term in the Sturm–Liouville ODE is positive because in the Sturm–Liouville equation

$$(py')' + qy + \lambda ry = 0$$

the eigenvalues λ are generally nonnegative. For instance, in Example 5 we anticipate (from the boundary conditions) that the expansion will be on the ϕ variable, so the Sturm–Liouville ODE is (76), not (75). Thus, we choose the $+\kappa^2$ in (74), so that the last term on the left side of (76) is $+\kappa^2\Phi$, not $-\kappa^2\Phi$.

EXERCISES 20.3

1. Show that if $u_1 = u_2 = 0$ and $f(\theta) = 100$, then (13), (15), and (17) give the result (19).

2. Solve for $u(r,\theta)$ and sketch, based on intuition, the $u = 25, 50, 75$ isotherms: $\nabla^2 u = 0$

(a) in $1 < r < 2$, $0 < \theta < \pi$; $u(r,0) = u(2,\theta) = 0$, $u(r,\pi) = u(1,\theta) = 100$
(b) in $1 < r < 2$, $0 < \theta < \pi$; $u_\theta(r,0) = u(2,\theta) = 0$, $u(r,\pi) = u(1,\theta) = 100$
(c) in $1 < r < 2$, $0 < \theta < \pi$; $u(r,\pi) = 100$, $u(r,0) = u_r(1,\theta) = u(2,\theta) = 0$
(d) in $1 < r < 2$, $-\infty < \theta < \infty$; $u(1,\theta) = 0$, $u(2,\theta) = 100$
(e) in $1 < r < 2, 0 < \theta < \pi$; $u(1,\theta) = u_\theta(r,0) = u_\theta(r,\pi) = 0$, $u(2,\theta) = 100$
(f) in $0 < r < 3, 0 < \theta < 3\pi/2$; $u(r,0) = u(r,3\pi/2) = 100$, $u(3,\theta) = 0$, u bounded
(g) in $0 < r < 3$, $0 < \theta < 3\pi/2$; $u_\theta(r,0) = u(3,\theta) = 0$, $u(r,3\pi/2) = 100$, u bounded
(h) in $0 < r < 3$, $0 < \theta < 3\pi/2$; $u(r,0) = u_r(3,\theta) = 0$, $u(r,3\pi/2) = 100$, u bounded
(i) in $0 < r < 3$, $0 < \theta < 3\pi/2$; $u(r,0) = u(r,3\pi/2) = 0$, $u(2,\theta) = 100$ on $0 < \theta < \pi/2$ and 0 on $\pi/2 < \theta < 3\pi/2$, u bounded
(j) in $3 < r < \infty$, $0 < \theta < \pi/2$; $u(r,0) = u(r,\pi/2) = 0$, $u(3,\theta) = 100$, u bounded
(k) in $3 < r < \infty$, $0 < \theta < \pi$; $u(r,0) = u_\theta(r,\pi) = 0$, $u(3,\theta) = 100$
(l) in $0 < r < 2$, $0 < \theta < \pi$; $u(r,0) = u(r,\pi) = 0$, $u(2,\theta) = 100$

3. Solve for $u(r,\theta)$ and give a labeled plot (by computer if necessary) of representative isotherms, as many as it takes to give a clear picture of the temperature field: $\nabla^2 u = 0$ in $r < 1$, u bounded, $u(1,0) = f(\theta)$.

(a) $f(\theta) = 50 + 20\cos\theta$
(b) $f(\theta) = 50 + 50(\cos\theta + \sin\theta)$
(c) $f(\theta) = 20\cos 2\theta$
(d) $f(\theta) = 20\sin 2\theta$
(e) $f(\theta) = 20\cos 3\theta$
(f) $f(\theta) = 20\sin 3\theta$
(g) $f(\theta) = 20\cos 4\theta$
(h) $f(\theta) = 20\cos 5\theta$

4. Consider a thin flat circular plate of radius b, that is thermally insulated on its two flat faces. With a hacksaw we make a radial cut along $\theta = 0$, say, from $r = b$ to $r = 0$. The small gap, due to the cut, may be approximated as a thermal insulator, so that $\partial u/\partial n = 0$ on the edges $\theta = 0$ and $\theta = 2\pi$. If the circumference of the plate is held at the temperature $50(1 + \sin\theta)$ for a long time, the steady-state temperature field

$u(r, \theta)$ is governed by the boundary-value problem

$$\begin{aligned} &\nabla^2 u = 0, \qquad (0 < r < b, \ \ 0 < \theta < 2\pi) \\ &\frac{\partial u}{\partial n}(r, 0) = \frac{\partial u}{\partial n}(r, 2\pi) = 0, \\ &u(b, \theta) = 50(1 + \sin\theta), \quad u \text{ bounded.} \end{aligned}$$

Solve for, and plot, the temperature distributions $u(r, 0)$ and $u(r, 2\pi)$ along the two edges of the cut. Does the answer depend in any way on whether the plate is steel or brass or whatever? Explain.

5. (*Plane with circular hole*) As a sort of "inverse" of Example 2, consider the domain to be the whole plane, with a circular *hole* of radius a:

$$\begin{aligned} &\nabla^2 u = 0, \qquad (a < r < \infty) \\ &u(a, \theta) = f(\theta), \quad u \text{ bounded as } r \to \infty. \end{aligned}$$

Solve for $u(r, \theta)$, leaving expansion coefficients in integral form. What is the value of u at $r = \infty$?

6. (*Plane flow over a circular bump*) First, read Example 3 of Section 16.10.

(a) Then, solve (38) in Section 16.10 and derive the solution

$$\Phi(r, \theta) = U\left(r + \frac{a^2}{r}\right)\cos\theta + C, \tag{6.1}$$

where C is an arbitrary constant that can be set equal to zero without loss. NOTE: The velocity field $\mathbf{v}$ is then available as $\mathbf{v} = \nabla\Phi$. Knowing $\mathbf{v}$, one could use the Bernoulli equation of fluid mechanics (which is derived in Exercise 12 of Section 16.10) to determine the pressure field and, in particular, the resulting aerodynamic force on the semicircular bump (which might, for instance, be the roof of a building). Observe that in this problem *a boundedness condition on* Φ *would be inappropriate* since $\Phi \sim Ur\cos\theta$ as $r \to \infty$. Nevertheless, the physical quantity $\mathbf{v} = \nabla\Phi$ is bounded: $\mathbf{v} \sim U\hat{\mathbf{i}}$ as $r \to \infty$. Finally, observe that (6.1) can be expressed as the superposition $\Phi = \Phi_1 + \Phi_2$, where $\Phi_1 = Ur\cos\theta = Ux$ is the potential of the "free stream," and $\Phi_2 = U(a^2/r)\cos\theta$ accounts for the disturbance caused by the presence of the bump (indeed, $\Phi_2 \to 0$ as $a \to 0$).

(b) By way of graphics, the most interesting display is not a display of constant Φ curves but a display of representative streamlines, as in Fig. 7 of Section 16.10. By streamlines we mean the constant Ψ curves, where Ψ is the stream function introduced in Exercise 8 of Section 16.10. From (8.3) therein, Ψ is related to Φ according to

$$\frac{\partial\Psi}{\partial x} = -\frac{\partial\Phi}{\partial y}, \qquad \frac{\partial\Psi}{\partial y} = \frac{\partial\Phi}{\partial x}. \tag{6.2a,b}$$

Use (6.2) and (6.1) to derive the result

$$\Psi(x, y) = U\left(y - \frac{a^2 y}{x^2 + y^2}\right). \tag{6.3}$$

(c) Then (with $U = a = 1$, say) use computer software such as the *Maple* implicitplot command to generate the streamline pattern that we sketched in Fig. 7 (Section 16.10). Choose the streamlines through the points $(x, y) = (-4, 0.2), (-4, 0.8), (-4, 1.4)$, and $(-4, 2)$, say.

7. (*Flow past a circular cylinder; nonuniqueness*) In Exercise 6 we consider the flow of a free stream past a semicircular bump. Here, we consider the flow of a free stream past a circular cylinder. The boundary-value problem is

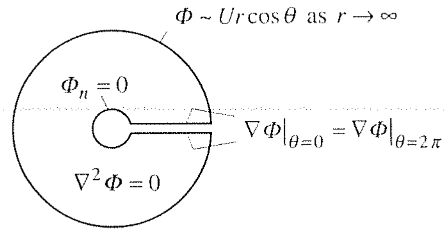
$$\begin{gathered} \nabla^2\Phi = \Phi_{rr} + \frac{1}{r}\Phi_r + \frac{1}{r^2}\Phi_{\theta\theta} = 0, \\ (a < r < \infty, \ \ 0 < \theta < 2\pi) \\ \Phi_n(a, \theta) = 0, \\ \nabla\Phi\Big|_{\theta=0} = \nabla\Phi\Big|_{\theta=2\pi} \quad \text{over } a < r < \infty, \\ \Phi \sim Ur\cos\theta \quad \text{as } r \to \infty. \end{gathered}$$

(7.1a,b,c,d)

That is, if we specify that $0 < \theta < 2\pi$, the radial lines $\theta = 0$ and $\theta = 2\pi$ ($a < r < \infty$) become part of the boundary (see the accompanying sketch), so that boundary conditions are needed along these lines. The appropriate boundary condition is that physical velocity $\mathbf{v}$ be the same at $\theta = 0$ and $\theta = 2\pi$ (for all $a < r < \infty$), which condition is expressed as (7.1c) or

$$\begin{aligned} \Phi_r(r, 0) &= \Phi_r(r, 2\pi), \qquad (a < r < \infty) \\ \frac{1}{r}\Phi_\theta(r, 0) &= \frac{1}{r}\Phi_\theta(r, 2\pi). \qquad (a < r < \infty) \end{aligned}$$

(7.1e,f)


$\Phi \sim Ur\cos\theta$ as $r \to \infty$
$\Phi_n = 0$
$\nabla\Phi|_{\theta=0} = \nabla\Phi|_{\theta=2\pi}$
$\nabla^2\Phi = 0$

Integrating (7.1e) with respect to r and cancelling r's in (7.1f), we obtain

$$\Phi(r,0) = \Phi(r,2\pi) + \Gamma, \quad \Phi_\theta(r,0) = \Phi_\theta(r,2\pi) \qquad (7.1\text{g,h})$$

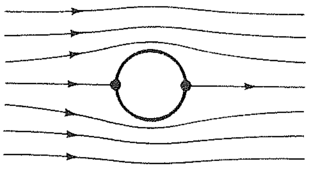
on $a < r < \infty$, where Γ is an arbitrary constant.

(a) Solve the resulting boundary-value problem (7.1a,b,g,h,d) and show that

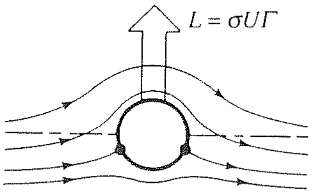
$$\Phi(r,\theta) = U\left(r + \frac{a^2}{r}\right)\cos\theta - \frac{\Gamma}{2\pi}\,\theta, \qquad (7.2)$$

which solution differs from (7.1) (with $C = 0$, say) only by the $-\Gamma\theta/2\pi$ term. NOTE: In physical terms, the $U(r + a^2r^{-1})\cos\theta$ term in (7.2) corresponds to a flow that is symmetric about the x axis (as sketched in the figure below), and which has *stagnation points* (i.e., where $\mathbf{v} = \mathbf{0}$) at $r = a$ and $\theta = 0, \pi$. The $-\Gamma\theta/2\pi$ term contributes the additional velocity

$$\mathbf{v} = \nabla\left(-\frac{\Gamma}{2\pi}\,\theta\right) = -\frac{\Gamma}{2\pi r}\hat{\mathbf{e}}_\theta,$$



which is a clockwise circular vortex flow (see Exercise 3 of Section 16.5) induced by a fictitious clockwise *vortex*, of strength Γ, at the origin – "fictitious" because there is no fluid inside the circle $r = a$. The vector superposition of these two velocity contributions gives a flow somewhat as we have sketched in the next figure, namely, the symmetric flow (the preceding figure) plus some clockwise "swirl" that is proportional to Γ.



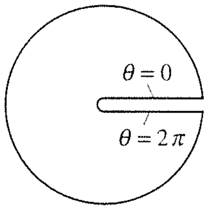
The two flows add on the upper part of the cylinder and subtract on the lower part, so there are higher velocities on the upper surface of the cylinder and lower velocities on the lower surface. Since Bernoulli's equation $\sigma v^2/2 + p =$ constant (σ = mass density, p = pressure) tells us that the higher the velocity the lower the pressure, and vice versa, it follows that a *lift force* is generated on the cylinder, $L = \sigma U\Gamma$ force per unit length of the cylinder (i.e., per unit z length).
(b) Show that the two stagnation points (in the preceding figure) are located on the cylinder surface by the equation $\sin\theta = -\Gamma/(4\pi Ua)$; e.g., if $\Gamma = 0$ then $\theta = \pi$ and 2π, as in the symmetric-flow figure. What happens regarding the existence and location of stagnation points if $\Gamma > 4\pi Ua$? Explain.
(c) From (7.2), $\mathbf{v} = \nabla\Phi$, and the Bernoulli equation, obtain the pressure distribution on the cylinder, integrate it, and thus derive the famous **Kutta–Joukowski lift formula**

$$L = \sigma U\Gamma \qquad (7.3)$$

stated above.

8. (*Alternative approach to Example 2*) In Example 2 we observe that $-\infty < \theta < \infty$ and impose the periodicity condition (28). Alternatively, we can consider that $0 < \theta < 2\pi$, in which case the lines $\theta = 0$ and $\theta = 2\pi$ occur as boundary edges of the domain. That is, we make an infinitely thin slit in the region as shown in the figure below. Since $0 < \theta < 2\pi$, we discard the 2π-periodicity condition, but we now have two artificially created boundaries, $\theta = 0$ and $\theta = 2\pi$, along which to specify boundary conditions. In particular, we impose the two conditions

$$\begin{aligned} u(r,0) &= u(r,2\pi), \\ u_\theta(r,0) &= u_\theta(r,2\pi) \end{aligned} \qquad (8.1,2)$$



over $0 < r < b$, so that both the temperature and heat flux are continuous across the slit. With the problem reformulated in this manner, solve for $u(r,\theta)$ and show that the solution obtained is the same as in Example 2. NOTE: The boundedness condition at $r = 0$ is still needed.

9. Derive the result

$$\frac{1}{2} + \sum_{n=1}^{\infty}\left(\frac{r}{b}\right)^n \cos n(\vartheta - \theta) = \frac{1}{2}\,\frac{b^2 - r^2}{b^2 - 2br\cos(\vartheta - \theta) + r^2} \qquad (9.1)$$

stated in Comment 2 of Example 2. HINT: Write

$$\sum_{n=1}^{\infty}\left(\frac{r}{b}\right)^{n}\cos n(\vartheta-\theta)=\operatorname{Re}\sum_{n=1}^{\infty}\left(\frac{r}{b}\right)^{n}e^{in(\vartheta-\theta)}, \tag{9.2}$$

and use the *geometric series*

$$\sum_{n=0}^{\infty}z^{n}=\frac{1}{1-z}, \tag{9.3}$$

which holds even if z is complex, provided that $|z|<1$. As usual, Re denotes the real part of the quantity.

10. Below Example 2 we show that the average value property of the Laplace equation holds in two dimensions, and at the end of Example 5 we showed that it holds in three dimensions. Here, we ask you to show that it holds for the *one*-dimensional Laplace equation $d^2u/dx^2=0$ as well.

11. Use the maximum and minimum principles to show, for the Laplace equation (in two or three dimensions), that if we change the boundary values only slightly, then the solution values (i.e., within the solution domain) change only slightly too. Specifically, show that if $\nabla^2 u_1=0$ in $\mathcal{D}$ with boundary values $u_1=f_1$, and $\nabla^2 u_2=0$ in $\mathcal{D}$ with boundary values $u_2=f_2$, then $\min(f_1-f_2)\le u_1-u_2\le \max(f_1-f_2)$.

12. (*Delta function behavior of Poisson kernel*) With $b=1$ and $\theta=1$, say, obtain computer plots of the Poisson kernel $P(r,\vartheta-\theta)$ versus ϑ, from $\vartheta=-\pi$ to $\vartheta=+\pi$, for these r values: $r=0,0.5,0.8$, and 0.9. Using *Maple*, for instance, you can use the plot command, which can be accessed by first typing the with(plots): command. NOTE: Surely, if the boundary temperature in (36) is $f(\vartheta)=\text{constant}=1$, then the solution will be $u(r,\theta)=\text{constant}=1$ as well. It follows that $\int_{-\pi}^{\pi}P(r;\vartheta-\theta)\,d\vartheta=1$ for all $0\le r<b$. Thus, each of your plots, for different r values, will have unit area. Further, they become increasingly focused at $\vartheta=\theta$ as $r\to b$. Thus, it appears that $P(r,\vartheta-\theta)$ is a *delta sequence* at θ as $r\to b$. In fact, the boundary condition requires of (36) that

$$f(\theta)=\lim_{r\to b}\int_{-\pi}^{\pi}P(r,\vartheta-\theta)f(\vartheta)\,d\vartheta,$$

which result confirms our suspicion: as $r\to b$, the Poisson kernel becomes a delta function at $\vartheta=\theta$, which picks out the value $f(\theta)$ and satisfies the boundary condition. This result is typical of linear PDE's: The solution due to Dirichlet boundary data f can be expressed as an integration, over the boundary, of a kernel times the boundary values. As a boundary point is approached from within the domain, the kernel becomes a delta function and picks out the value of f at that boundary point, thereby satisfying the boundary condition.

13. Consider the Dirichlet problem

$$\nabla^2 u=u_{rr}+\frac{1}{r}u_r+\frac{1}{r^2}u_{\theta\theta}=0 \tag{13.1}$$

in $a<r<b$, $0<\theta<\alpha$, with boundary conditions

$$\begin{gathered}u(r,0)=u(a,\theta)=u(b,\theta)=0,\\ u(r,\alpha)=f(r).\end{gathered} \tag{13.2,3}$$

(a) Seeking $u(r,\theta)=R(r)\Theta(\theta)$ and anticipating the Fourier expansion along the $\theta=\alpha$ edge, obtain

$$\begin{aligned}u(r,\theta)&=(A+B\ln r)(E+F\theta)\\ &+[C\cos(\kappa\ln r)+D\sin(\kappa\ln r)](G\cosh\kappa\theta+H\sinh\kappa\theta).\end{aligned}$$

(b) Applying the boundary conditions (13.2), arrive at

$$u(r,\theta)=\sum_{n=1}^{\infty}I_n\sin\left(\kappa_n\ln\frac{r}{a}\right)\sinh\kappa_n\theta, \tag{13.4}$$

where

$$\kappa_n=n\pi/\ln\left(\frac{b}{a}\right). \tag{13.5}$$

(c) Applying the boundary condition (13.3), show that

$$\begin{aligned}I_n&=\frac{1}{\sinh\kappa_n\alpha}\,\frac{\displaystyle\int_a^b f(r)\phi_n(r)\frac{1}{r}\,dr}{\displaystyle\int_a^b\phi_n^2(r)\frac{1}{r}\,dr}\\ &=\frac{2}{\ln\left(\dfrac{b}{a}\right)\sinh\kappa_n\alpha}\int_a^b f(r)\phi_n(r)\frac{1}{r}\,dr,\end{aligned} \tag{13.6}$$

where $\phi_n(r)=\sinh[\kappa_n\ln(r/a)]$. HINT: The ϕ_n's are the eigenfunctions of the Sturm – Liouville problem on $R(r)$. Identify that problem (i.e., the ODE, the r interval, the boundary conditions on R, and the weight function).

EXERCISES FOR THE OPTIONAL SECTIONS 20.3.2, 20.3.3

14. Derive the $\kappa\ne0$ part of the general solution (43) using equation (50) in Section 4.6.

15. Derive the $\kappa\ne0$ part of the general solution (57) using equation (50) in Section 4.6.

16. Solve by separation of variables, leaving expansion coefficients in integral form; u is to be bounded, and

$$\nabla^2 u=u_{rr}+\frac{1}{r}u_r+u_{zz}=0$$

(a) in $0 \leq r < b$, $0 < z < \infty$, with $u(b,z) = 0$, $u(r,0) = f(r)$
(b) in $0 \leq r < b$, $-\infty < z < \infty$, with $u(b,z) = f(z)$, where $f(z)$ is a $2L$-periodic square wave defined over one period as 100 over $0 < z < L$ and 0 over $L < z < 2L$
(c) in $0 \leq r < b$, $0 < z < L$, with $u_z(r,0) = u_z(r,L) = 0$, $u(b,z) = 50$
(d) in $0 \leq r < b$, $0 < z < L$, with $u_z(r,0) = u(r,L) = 0$, $u(b,z) = 50$
(e) in $0 \leq r < b$, $0 < z < L$, with $u_z(r,0) = f(r)$, $u(r,L) = u(b,z) = 0$
(f) in $a \leq r < \infty$, $0 < z < L$, with $u(r,0) = u_1$, $u(r,L) = u_2$, $u(a,z) = u_3$
(g) in $a \leq r < \infty$, $0 < z < \infty$, with $u(r,0) = 0$, $u(a,z) = 25\sin(3z/2)$

17. Show that the change of variables $\mu = \cos\phi$ does, indeed, change (76) to the Legendre equation (77).

18. In Example 5, let $f(\phi)$ be 0 on the bottom half of the sphere ($\pi/2 < \phi < \pi$, $-1 < \mu < 0$) and 100 on the top half ($0 < \phi < \pi/2$, $0 < \mu < 1$). Evaluating the first several A_n's, show that (81) becomes

$$u(\rho,\phi) = 50\left[P_0(\cos\phi) + \frac{3}{2}\left(\frac{\rho}{c}\right)P_1(\cos\phi) - \frac{7}{8}\left(\frac{\rho}{c}\right)^3 P_3(\cos\phi) + \frac{11}{16}\left(\frac{\rho}{c}\right)^5 P_5(\cos\phi) - \frac{75}{128}\left(\frac{\rho}{c}\right)^7 P_7(\cos\phi) + \cdots\right].$$

HINT: It is simplest to evaluate the integral $\int_{-1}^{1} fP_n\,d\mu = 100\int_0^1 P_n(\mu)\,d\mu$ using computer software. Using *Maple*, for example, first enter with(orthopoly): to access the Legendre polynomials. Then, with $n = 5$ say, enter int$(P(5,x), x = 0..1)$; to evaluate $\int_0^1 P_5(\mu)\,d\mu$.

19. (*Variations on Example 5*) Solve the Laplace equation (71) in spherical coordinates, with symmetry about the z axis, and evaluate the expansion coefficients so as to obtain the first five (if there are that many) nonvanishing terms of the series solutions, as we did in Exercise 18:

(a) in $0 \leq \rho < c$, $0 < \phi < \pi/2$, with $u(\rho,\pi/2) = 0$, $u(c,\phi) = 100$
(b) in $0 \leq \rho < c$, $0 < \phi < \pi/2$, with (the normal derivative) $u_n(\rho,\pi/2) = 0$, $u(c,\phi) = 100$
(c) in $c < \rho < \infty$, $0 < \phi < \pi/2$, with $u(\rho,\pi/2) = 0$, $u(c,\phi) = 100$
(d) in $c < \rho < \infty$, $0 < \phi < \pi/2$, with $u_\rho(c,\phi) = 0$, $u(\rho,\pi/2) = 100$

20.4 Fourier Transform (Optional)

In Section 18.4 we studied the solution of diffusion problems by the Fourier and Laplace transforms. Laplace transforming on the t variable is always an option for the diffusion equation because $0 < t < \infty$, and the problem is of initial-value type with respect to t. However, it is of boundary-value type with respect to x (i.e., there is a boundary condition at each end) so, alternatively, we can use a Fourier transform on x if the x domain is $-\infty < x < \infty$ or a Fourier cosine or sine transform on x if the domain is $0 < x < \infty$.

In contrast, the Laplace equation is of boundary-value type in both independent variables, so the Laplace transform is not helpful. Still, if the domain is infinite in one of the independent variables then we can employ a Fourier transform on that variable; if it is semi-infinite then we can employ a Fourier cosine or sine transform.

EXAMPLE 1. *Dirichlet Problem for Half Plane.* Consider the half-plane problem

$$\nabla^2 u = u_{xx} + u_{yy} = 0, \qquad (-\infty < x < \infty,\ 0 < y < \infty) \tag{1a}$$

$$u(x,0) = f(x) \qquad (-\infty < x < \infty) \tag{1b}$$

depicted in Fig. 1, realizing that we may be stipulating additional boundary conditions as we proceed.

Fourier transform (1a) with respect to x:

$$F\{u_{xx} + u_{yy}\} = F\{0\}, \tag{2a}$$

$$F\{u_{xx}\} + F\{u_{yy}\} = 0, \tag{2b}$$

$$(i\omega)^2 \hat{u} + \int_{-\infty}^{\infty} \frac{\partial^2 u}{\partial y^2} e^{-i\omega x}\, dx = 0, \tag{2c}$$

$$-\omega^2 \hat{u} + \frac{d^2}{dy^2} \int_{-\infty}^{\infty} u(x,y) e^{-i\omega x}\, dx = 0, \tag{2d}$$

or

$$\frac{d^2\hat{u}}{dy^2} - \omega^2 \hat{u} = 0, \tag{3}$$

with general solution

$$\hat{u}(\omega, y) = Ae^{|\omega| y} + Be^{-|\omega| y}. \tag{4}$$

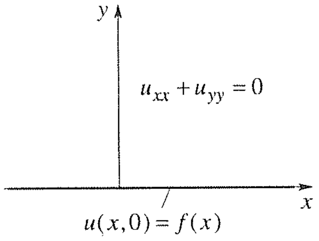


Figure 1. Problem (1).

(Let us defer discussion of the absolute value signs to Comment 1 below.) Recall from our study of the Fourier transform that for $F\{u_{xx}\}$ to equal $(i\omega)^2\hat{u}$ we need

$$u \to 0 \text{ and } u_x \to 0 \text{ as } x \to \pm\infty, \tag{5}$$

so let us suppose that u does satisfy the boundary conditions (5) to the east ($x \to +\infty$) and to the west ($x \to -\infty$). Condition (1b) is our southern boundary condition, but we are still lacking a second y boundary condition, to the north as $y \to \infty$. If we assume that

$$u(x,y) \to 0 \text{ as } y \to \infty, \tag{6}$$

then we formally obtain

$$\begin{aligned} \lim_{y\to\infty} \hat{u}(\omega, y) &= \lim_{y\to\infty} \int_{-\infty}^{\infty} u(x,y) e^{-i\omega x}\, dx \\ &= \int_{-\infty}^{\infty} \left[\lim_{y\to\infty} u(x,y)\right] e^{-i\omega x}\, dx \\ &= \int_{-\infty}^{\infty} 0\, e^{-i\omega x}\, dx = 0. \end{aligned} \tag{7}$$

Applying the latter result to (4) reveals that we need $A = 0$, so

$$\hat{u}(\omega, y) = Be^{-|\omega| y}. \tag{8}$$

To evaluate B take the transform of (1b),

$$\hat{u}\Big|_{y=0} = \hat{f}(\omega), \tag{9}$$

and impose that condition on (8):

$$\hat{u}\Big|_{y=0} = \hat{f}(\omega) = B. \tag{10}$$

Thus,

$$\hat{u}(\omega, y) = \hat{f}(\omega)e^{-|\omega|y}.$$

If we use entry 1 in Appendix D, along with the Fourier convolution property (entry 21), we obtain the final result

$$u(x, y) = f(x) * \frac{y}{\pi}\frac{1}{x^2 + y^2}, \tag{11}$$

or

$$\boxed{\begin{aligned} u(x, y) &= \frac{y}{\pi}\int_{-\infty}^{\infty} \frac{f(\xi)}{(x-\xi)^2 + y^2}\, d\xi \\ &\equiv \int_{-\infty}^{\infty} P(\xi - x, y) f(\xi)\, d\xi; \end{aligned}} \tag{12}$$

(12) is the **Poisson integral formula for the half plane**, and $P(\xi - x, y)$ is the corresponding **Poisson kernel**. The analogous formula for the circular disk is given in (36) of Section 20.3.

COMMENT 1. Why did we express the solution of (3) as $Ae^{|\omega|y} + Be^{-|\omega|y}$ rather than as

$$\hat{u}(\omega, y) = Ce^{\omega y} + De^{-\omega y}? \tag{13}$$

The forms (4) and (13) are indeed equivalent, but (4) is more convenient for applying the northern boundary condition (7) (namely, that $\hat{u} \to 0$ as $y \to \infty$). For remember that the Fourier inversion formula involves an integral on ω from $\omega = -\infty$ to $\omega = +\infty$. Thus, we need to allow for $\omega > 0$ and $\omega < 0$ in (13) and conclude, from (7), that $C(\omega) = 0$ for $\omega > 0$ and $D(\omega) = 0$ for $\omega < 0$, which story is more complicated than observing, in (4), that $e^{|\omega|y}$ is the "bad" term and $e^{-|\omega|y}$ is the "good" term so that we need to set $A = 0$.

COMMENT 2. Let us focus our attention on the kernel P. Observe that P has unit area, for each $y > 0$, since

$$\int_{-\infty}^{\infty} P(\xi - x, y)\, d\xi = \frac{y}{\pi}\int_{-\infty}^{\infty} \frac{d\xi}{(\xi - x)^2 + y^2} = 1, \tag{14}$$

and that its graph becomes more and more sharply focused (at $\xi = x$) as $y \to 0$, as seen in Fig. 2. Thus, $P(\xi - x, y)$ looks like a delta sequence at $\xi = x$, as indeed must be true since the boundary condition (1b) really means that $\lim_{y\to 0} u(x, y) = f(x)$ or, since $u(x, y)$ is given by (12),

$$\lim_{y\to 0}\int_{-\infty}^{\infty} P(\xi - x, y) f(\xi)\, d\xi = f(x), \tag{15}$$

which, by definition, means that P becomes a delta function at x as $y \to 0$.

COMMENT 3. As a check case, let us use (12) for the case where $u(x, 0) = f(x) =$ constant $= f_0$, since then the solution should, by inspection, be $u(x, y) = f_0$ everywhere. In fact, (12) does give that correct result – even though the assumed conditions at infinity ($u \to 0$ as $x \to \pm\infty$ and as $y \to \infty$) are not satisfied. That is, (12) is even more robust than anticipated. ▮

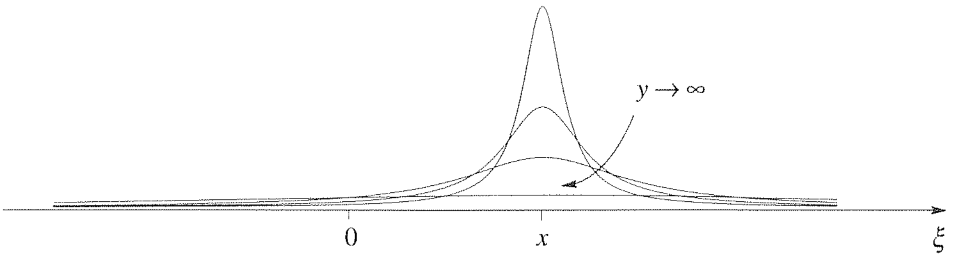


Figure 2. Poisson kernel P.

Closure. Besides illustrating the use of the Fourier transform in solving the Laplace equation on an infinite domain, we also obtain an important specific solution, the Poisson integral formula (12) for the half plane. As usual (see Exercise 12 of Section 20.3), the solution due to Dirichlet boundary data f can be expressed as an integration, over the boundary, of a kernel times the boundary values. As a boundary point is approached from within the domain, the kernel becomes a delta function and picks out the value of f at that point, thereby satisfying the boundary condition.

EXERCISES 20.4

1. (a) Use (12) to evaluate $u(x,y)$ if $f(x) = 100H(x)$ (where H is the Heaviside function).
(b) Draw the isotherms $u = 25, 50, 75$.
(c) Plot $u(x,y)$ versus x at $y = 0, 1, 3$.

2. (a) Use (12) to evaluate $u(x,y)$ if $f(x) = 100[H(x+1) - H(x-1)]$.
(b) Plot $u(x,y)$ versus x at $y = 0$ and at $y = 2$.
(c) Show that $u(x,y) \sim 200/(\pi y)$ as $y \to \infty$. HINT: Note that

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

for $|x| < 1$, and

$$\tan^{-1} x = \frac{\pi}{2} - \frac{1}{x} + \frac{1}{3x^3} - \frac{1}{5x^5} + \cdots$$

for $|x| < 1$ where, in each case, $\tan^{-1}$ denotes the choice (of the multivalued $\tan^{-1}$ function) lying between $-\pi/2$ and $+\pi/2$.

3. (a) Show, from (12), that if $f(x)$ is an even function of x then so is $u(x,y)$.
(b) Show, from (12), that if $f(x)$ is an odd function of x then so is $u(x,y)$.

4. Use (12) and the method of images (explained in the optional Section 18.5) to solve $\nabla^2 u = u_{xx} + u_{yy} = 0$ in the first quadrant ($x > 0, y > 0$), with $u(x,0) = f(x)$, with suitable conditions at $x = \infty$ and at $y = \infty$, and with

(a) $u(0,y) = 0$ (b) $u_x(0,y) = 0$

5. Consider the infinite strip problem

$$\nabla^2 u = u_{xx} + u_{yy} = 0, \qquad (|x| < \infty, \ 0 < y < a)$$
$$u(x,0) = f(x), \quad u(x,a) = g(x).$$

(a) Show that

$$u(x,y) = F^{-1}\left\{\hat{f}(\omega)\,\frac{\sinh \omega(a-y)}{\sinh \omega a} + \hat{g}(\omega)\,\frac{\sinh \omega y}{\sinh \omega a}\right\},$$

but do not try to evaluate that Fourier inverse.
(b) With $f(x) = g(x) = 100H(x)$, use intuition to sketch the isotherms, say $u = 10, 25, 50, 75, 90$.

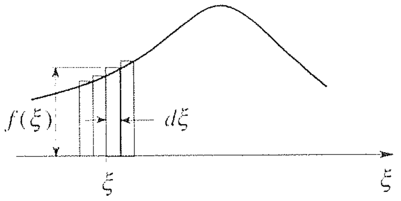
6. In Comment 2 we discuss the delta function behavior of the Poisson kernel P as $y \to 0$. In fact, letting $f(x) = \delta(x - x_0)$, show that $P(x_0 - x, y) = P(x - x_0, y)$ is itself the solution or "response" due to a boundary temperature that is a delta function at x_0. NOTE: With this result in mind, we can interpret (12) as a superposition principle. For let us break f into narrow vertical rectangles. What is the response $du(x,y)$ due to the single shaded rectangular pulse (shown in the figure)? The pulse is a delta function at ξ (as $d\xi \to 0$), so (as noted above) its response at x, y is $P(\xi - x, y)$; actually, the pulse is a delta function *scaled* by $f(\xi)\,d\xi$ because its area is $f(\xi)\,d\xi$ rather than unity, so its response is likewise scaled,

$$du(x,y) = P(x-\xi, y) f(\xi)\,d\xi$$

or, since P is an even function of its first argument,

$$du(x,y) = P(\xi - x, y) f(\xi)\,d\xi. \tag{6.1}$$

And superimposing these infinitesimal responses (by integration) gives (12). Thus, we can understand (12) as a *superposition principle.*



20.5 Numerical Solution

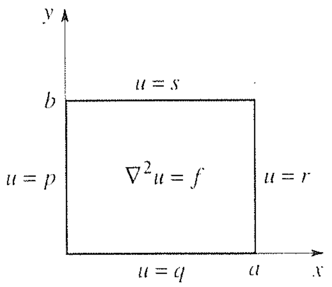


Figure 1. The problem (1).

20.5.1. Rectangular domains. Following the same general lines as in Section 18.6, where we develop the finite-difference solution technique for the diffusion equation, here we do the same for the Laplace equation or, more generally and with no additional difficulty, for the Poisson equation $\nabla^2 u = f$.

Limiting our attention to the two-dimensional case, we begin with the problem

$$\nabla^2 u = u_{xx} + u_{yy} = f(x,y), \qquad (0 < x < a,\ 0 < y < b) \tag{1a}$$

$$u(0,y) = p(y),\ \ u(x,0) = q(x),\ \ u(a,y) = r(y),\ \ u(x,b) = s(x) \tag{1b}$$

depicted in Fig. 1, and generalize to nonrectangular domains in Section 20.5.2. Seeking an approximate numerical solution, we discretize the problem by dividing a into M equal parts of dimension $\Delta x = a/M$, dividing b into N equal parts of dimension $\Delta y = b/N$, and defining nodal points $P_{jk} = (x_j, y_k) = (j\Delta x, k\Delta y)$ for $j = 0, 1, 2, \ldots, M$ and $k = 0, 1, 2, \ldots, N$. Accordingly, we seek u not everywhere in the domain but only at the nodal points – more specifically, at the interior nodal points since u is prescribed at the boundary nodal points by the Dirichlet boundary conditions (1b). The grid is shown in Fig. 2 for the choice $M = N = 3$.

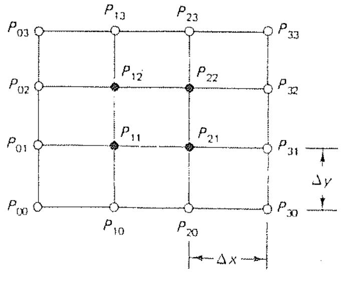


Figure 2. Finite-difference mesh for $M = N = 3$.

Next, we replace the PDE (1a) by a finite-difference approximation that will lead to a set of linear algebraic equations in the unknown nodal values of u. As in Section 18.6, we adopt the approximations

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{u(x-\Delta x, y) - 2u(x,y) + u(x+\Delta x, y)}{(\Delta x)^2}, \tag{2a}$$

$$\frac{\partial^2 u}{\partial y^2} \approx \frac{u(x, y-\Delta y) - 2u(x,y) + u(x, y+\Delta y)}{(\Delta y)^2}, \tag{2b}$$

Putting (2a, b) into the PDE, with $x = x_j$, $x - \Delta x = x_{j-1}$, $x + \Delta x = x_{j+1}$ (and similarly for y) gives

$$\frac{u(x_{j-1}, y_k) - 2u(x_j, y_k) + u(x_{j+1}, y_k)}{(\Delta x)^2} + \frac{u(x_j, y_{k-1}) - 2u(x_j, y_k) + u(x_j, y_{k+1})}{(\Delta y)^2} \approx f(x_j, y_k). \qquad (3)$$

Thus, we adopt the algebraic equation

$$\boxed{\frac{U_{j-1,k} - 2U_{j,k} + U_{j+1,k}}{(\Delta x)^2} + \frac{U_{j,k-1} - 2U_{j,k} + U_{j,k+1}}{(\Delta y)^2} = f_{j,k}} \qquad (4)$$

as our finite-difference approximation of (1a), where $f_{j,k}$ is shorthand for $f(x_j, y_k)$. As in Section 18.6, we use different letters (lowercase and uppercase) to distinguish between the exact solution $u(x, y)$ of (1) and the approximating solution $U_{j,k}$ generated by the finite-difference equation (4). We call $u(x_j, y_k) - U_{j,k}$ the **truncation error** at $P_{j,k}$, namely, the error incurred by replacing u_{xx} and u_{yy} in (1a) by the finite-difference approximations (2). Observe that whereas in Section 18.6 we distinguish between the *local* truncation error (incurred in carrying out a single time step) and the *accumulated* truncation error (incurred in carrying out all the time steps up until the time in question) – here we do not – because *there are no time steps*. Thus, there is simply "the truncation error."

Suppose that we compute $U_{j,k}$ at a particular point P in the domain, then again using a finer mesh, again using a finer mesh, and so on. If, as the mesh becomes infinitely fine (i.e., as Δx and Δy both tend to zero) the computed values converge to the exact solution at P, then the finite-difference scheme is **convergent**. Recall that in our study of the diffusion equation (Section 18.6) we pay comparable attention to the companion questions of convergence and stability; the difference scheme is said to be stable if the accumulated roundoff error remained small. For the Poisson and Laplace equations, however, we do not "march out" a solution in time, so the issue of stability is not relevant. In fact, roundoff error should be quite negligible compared with the truncation error for the methods considered in this section. If we choose $\Delta x = \Delta y \equiv h$, say, then (4) becomes

$$\boxed{U_{j-1,k} + U_{j,k-1} + U_{j+1,k} + U_{j,k+1} - 4U_{j,k} = h^2 f_{j,k},} \qquad (5)$$

which is often expressed, schematically, in the form

$$\begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix} U = h^2 f.$$

If, in addition, $f(x, y) = 0$, so that (1a) reduces to the Laplace equation, then (5) gives

$$U_{j,k} = \frac{1}{4}\left(U_{j-1,k} + U_{j,k-1} + U_{j+1,k} + U_{j,k+1}\right). \qquad (6)$$

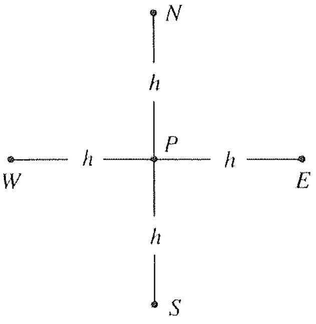


Figure 3. The five points.

We call these **five-point formulas** because they involve five grid points, which we denote as P, W(est), S(outh), E(ast), and N(orth), respectively, in Fig. 3. It is striking that (6) is a discrete and approximate version of the average value property of the two-dimensional Laplace equation, namely, that u at any given point P in the solution domain is the average value of u over any circle centered at P and lying entirely within the domain.

EXAMPLE 1. To illustrate the use of (5), let $a = b = 1$ and $f(x, y) = p(y) = q(x) = s(x) = 0$, and $r(y) = 100 \sin \pi y$ so that, by separation of variables, we have the simple exact solution

$$u(x, y) = 100 \frac{\sinh \pi x}{\sinh \pi} \sin \pi y \tag{7}$$

available for comparison with our computed approximate solution. As the simplest (and crudest) case, let $M = N = 2$. Then $\Delta x = \Delta y = 0.5$ and there is only one internal node, (Fig. 4). Writing out (5) for that point (i.e., with $j = k = 1$) gives

$$P_{11}: \quad U_{01} + U_{10} + U_{21} + U_{12} - 4U_{11} = 0 \tag{8}$$

or, recalling the given boundary conditions, $0 + 0 + 100 \sin (\pi/2) - 4U_{11} = 0$. Solving, $U_{11} = 25$ and we have the following comparison.

$$\begin{aligned} &\text{Computed:} && U_{11} = 25 && \text{(9a)} \\ &\text{Exact:} && u_{11} = 100 \frac{\sinh (\pi/2)}{\sinh \pi} \sin \frac{\pi}{2} = 19.9, && \text{(9b)} \end{aligned}$$

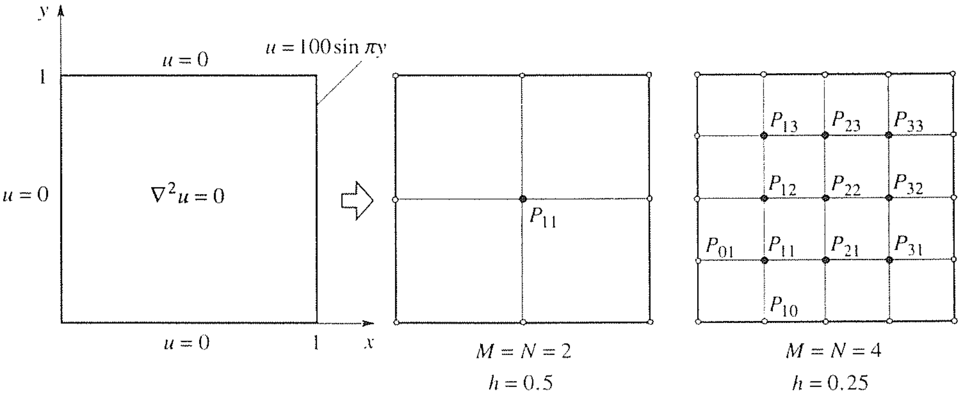


Figure 4. Example 1.

where u_{jk} means the exact solution $u(x, y)$ evaluated at P_{jk}; note that we will generally omit the comma between the two subscripted indices, for brevity. It is not surprising that the error is so great because the grid is so coarse; that is, $h = 0.5$ is not small compared to $a = b = 1$.

Next, let $M = N = 4$ (right-hand member of Fig. 4). Write out (5) for the nine internal grid points $P_{11}, P_{21}, \ldots, P_{23}, P_{33}$:

$$
\begin{array}{lrcl}
P_{11} & 0+0+U_{21}+U_{12}-4U_{11} & = & 0, \\
P_{21} & U_{11}+0+U_{31}+U_{22}-4U_{21} & = & 0, \\
 & & \vdots & \\
P_{23} & U_{13}+U_{22}+U_{33}+0-4U_{23} & = & 0, \\
P_{33} & U_{23}+U_{32}+100\sin\dfrac{3\pi}{4}+0-4U_{33} & = & 0
\end{array}
\tag{10}
$$

or, in matrix form,

$$
\left[\begin{array}{ccc|ccc|ccc}
-4 & 1 & 0 & 1 & & & & \cdots & 0 \\
1 & -4 & 1 & 0 & 1 & & & & \vdots \\
0 & 1 & -4 & 0 & 0 & 1 & & & \\
\hline
1 & 0 & 0 & -4 & 1 & 0 & 1 & & \\
 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & \\
 & & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\
\hline
 & & & 1 & 0 & 0 & -4 & 1 & 0 \\
\vdots & & & & 1 & 0 & 1 & -4 & 1 \\
0 & \cdots & & & & 1 & 0 & 1 & -4
\end{array}\right]
\left[\begin{array}{c}
U_{11} \\ U_{21} \\ U_{31} \\ \hline U_{12} \\ U_{22} \\ U_{32} \\ \hline U_{13} \\ U_{23} \\ U_{33}
\end{array}\right]
=
\left[\begin{array}{c}
0 \\ 0 \\ -100\sin\frac{\pi}{4} \\ \hline 0 \\ 0 \\ -100\sin\frac{\pi}{2} \\ \hline 0 \\ 0 \\ -100\sin\frac{3\pi}{4}
\end{array}\right],
\tag{11}
$$

where matrix elements not shown are zeros and the partitioning lines are to be ignored for the moment. Roughly speaking, the first equation in (10) ensures the satisfaction of the Laplace equation in the neighborhood of P_{11}, the second equation in (10) ensures the satisfaction of the Laplace equation in the neighborhood of P_{21}, and so on, so that the satisfaction of (10) is equivalent to the approximate satisfaction of the Laplace equation in the entire domain (as well as the Dirichlet boundary conditions).

Solving (11) by a computer algebra system (e.g., using the *Maple* linsolve command described at the end of Section 8.3), we obtain these values.

$$
\begin{array}{llll}
\text{Computed:} & U_{11}=5.8, & U_{21}=15.1, & U_{31}=33.2, \\
 & U_{12}=8.3, & U_{22}=21.3, & U_{32}=46.9, \\
 & U_{13}=5.8, & U_{23}=15.1, & U_{33}=33.2
\end{array}
\tag{12a}
$$

$$
\begin{array}{llll}
\text{Exact:} & u_{11}=5.3, & u_{21}=14.1, & u_{31}=32.0, \\
 & u_{12}=7.5, & u_{22}=19.9, & u_{32}=45.3, \\
 & u_{13}=5.3, & u_{23}=14.1, & u_{33}=32.0.
\end{array}
\tag{12b}
$$

These results are seen to be in better agreement than (9a) and (9b) but are still quite crude. If this were a realistic application, rather than only an illustration of the method, we might choose h to be 0.05 or smaller.

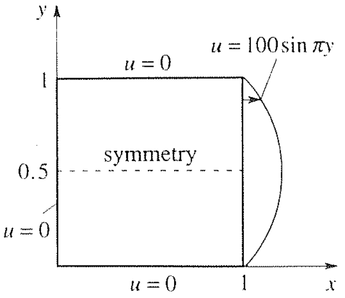


Figure 5. Symmetry.

COMMENT 1. Observe from Fig. 5 that both the domain and the boundary conditions are symmetric about the mid-line $y = 0.5$, so it is evident that the solution $u(x, y)$ should, likewise, be symmetric about that line.* Thus, our numerical solution, for the case where $M = N = 4$, is wasteful because we know in advance that

$$U_{13} = U_{11}, \quad U_{23} = U_{21}, \quad U_{33} = U_{31}, \tag{13}$$

so there are really only six unknowns rather than nine. It suffices to apply (5) to the six points $P_{11}, P_{21}, P_{31}, P_{12}, P_{22}, P_{33}$, and to use (13). Thus, the reduced system is as follows:

$$\begin{aligned}
P_{11}: &\quad 0 + 0 + U_{21} + U_{12} - 4U_{11} = 0, \\
P_{21}: &\quad U_{11} + 0 + U_{31} + U_{22} - 4U_{21} = 0, \\
P_{31}: &\quad U_{21} + 0 + 100\sin\tfrac{\pi}{4} + U_{32} - 4U_{31} = 0, \\
P_{12}: &\quad 0 + U_{11} + U_{22} + \underline{U}_{11} - 4U_{12} = 0, \\
P_{22}: &\quad U_{12} + U_{21} + U_{32} + \underline{U}_{21} - 4U_{22} = 0, \\
P_{32}: &\quad U_{22} + U_{31} + 100\sin\tfrac{\pi}{2} + \underline{U}_{31} - 4U_{32} = 0,
\end{aligned} \tag{14}$$

where the underlined terms are those that result from the symmetry relations (13).

COMMENT 2. Observe that (11) may be partitioned, according to the thin lines in (11), as

$$\begin{bmatrix} \mathbf{B} & \mathbf{I} & & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{B} & \mathbf{I} & & \vdots \\ & & \ddots & & \\ \vdots & & \mathbf{I} & \mathbf{B} & \mathbf{I} \\ \mathbf{0} & \cdots & & \mathbf{I} & \mathbf{B} \end{bmatrix} \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \vdots \\ \mathbf{U}_{N-2} \\ \mathbf{U}_{N-1} \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_{N-2} \\ \mathbf{c}_{N-1} \end{bmatrix}, \tag{15}$$

where $\mathbf{B}$ is the $(N-1) \times (N-1)$ matrix

$$\mathbf{B} = \begin{bmatrix} -4 & 1 & & \cdots & 0 \\ 1 & -4 & 1 & & \vdots \\ & & \ddots & & \\ \vdots & & 1 & -4 & 1 \\ 0 & \cdots & & 1 & -4 \end{bmatrix}, \tag{16}$$

and $\mathbf{I}$ is an identity matrix of order $N-1$. In (11), for instance, $N = 4$ so $\mathbf{B}$ and $\mathbf{I}$ are 3×3. Each vector is comprised of the unknown nodal values across the jth row of the mesh. Whereas $\mathbf{B}$ is tridiagonal, $\mathbf{A}$ is not tridiagonal; it is **block** tridiagonal. ■

The method is powerful because it enables us to obtain solutions even if the inputs $p(y), q(x), r(y), s(x)$, and $f(x, y)$ are nonconstant functions, in which case

*It surely seems clear, if only intuitively, that the solution is symmetric about the line $y = 0.5$ as claimed, but to put that claim on solid ground we can put forward arguments similar to those given in Section 18.5 on the method of images. We will leave that point for the exercises.

analytical solution becomes extremely laborious. However, it is to be appreciated that the computer calculation is not trivial if we seek good accuracy because (in the absence of helpful symmetries) we need to solve a system of $(M-1)\times(N-1)$ linear algebraic equations. If, for instance, we choose $M=N=50$, for the sake of accuracy, then we have a system of 2,401 equations! In such cases one concern is how to solve that large system of equations *efficiently*, to which topic we return in Section 20.5.3. First, in Section 20.5.2, we indicate how to extend the method to handle domains of essentially arbitrary shape.

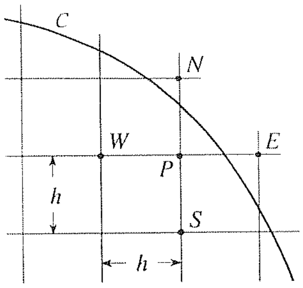


Figure 6. Nonrectangular domain.

20.5.2. Nonrectangular domains. Thus far we have dealt with cases where the boundary curve is rectangular so that grid lines can coincide with the edges of the domain as in Fig. 4, that is, where the mesh "fits" the domain. What happens if the mesh does not fit, as illustrated in Fig. 6? We cannot apply the finite-difference scheme (5) at grid points such as P because the points N and E do not fall on the boundary curve C; they fall outside the domain. To handle this case we slide N and E so that they do fall on C, as shown in Fig. 7, and revise the difference quotient approximations (2a,b) accordingly. [We need to modify (2a), for instance, because it gives u_{xx} at P as a linear combination of the values of u at W, P, and E. If we move E, then the weighting of those values will change; we can expect U_E to be weighted more heavily than U_W because E is closer to P than W. Similarly for (2b).] For the geometry shown in Fig. 7 we need to slide N and E, but in other cases we may need to slide W and/or S as well, so let us consider the most general case shown in Fig. 8, where $0<\alpha\leq 1$, $0<\beta\leq 1$, $0<\gamma\leq 1$, and $0<\delta\leq 1$.

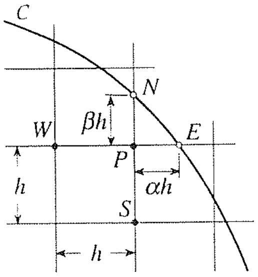


Figure 7. Adjusting the mesh near C.

We begin with Taylor expansions about P (namely, the point x_j, y_k) in the eastern and western directions, respectively,

$$u(x_j+\alpha h, y_k) = u(x_j, y_k) + u_x(x_j, y_k)\alpha h + \frac{1}{2!}u_{xx}(x_j, y_k)(\alpha h)^2 + \cdots,$$

$$u(x_j-\gamma h, y_k) = u(x_j, y_k) + u_x(x_j, y_k)(-\gamma h) + \frac{1}{2!}u_{xx}(x_j, y_k)(-\delta h)^2 + \cdots,$$

or, using N, E, S, W, P subscript notation instead,

$$u_E = u_P + u_x|_P\,\alpha h + \frac{1}{2}\,u_{xx}|_P\,\alpha^2 h^2 + \cdots, \tag{17a}$$

$$u_W = u_P - u_x|_P\,\gamma h + \frac{1}{2}\,u_{xx}|_P\,\gamma^2 h^2 - \cdots. \tag{17b}$$

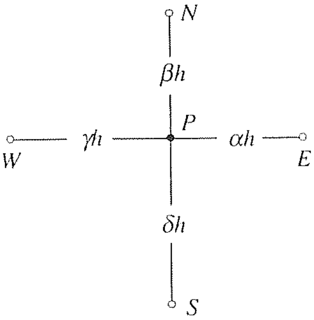


Figure 8. Most general case.

Multiplying (17a) by γ and (17b) by α and adding, to cancel the u_x terms, gives

$$\gamma u_E + \alpha u_W = (\gamma+\alpha)u_P + \frac{1}{2}(\alpha^2\gamma + \alpha\gamma^2)h^2\,u_{xx}|_P + \cdots, \tag{18}$$

so that if we neglect terms of order h^3 and higher in (18) then we obtain

$$u_{xx}|_P \approx \frac{2}{\alpha\gamma(\alpha+\gamma)h^2}\left[\gamma u_E + \alpha u_W - (\gamma+\alpha)u_P\right] \tag{19a}$$

in place of (2a). Similarly, Taylor expansions about P in the northern and southern directions give the result

$$u_{yy}\big|_P \approx \frac{2}{\beta\delta(\beta+\delta)h^2}\left[\delta u_N + \beta u_S - (\delta+\beta)u_P\right] \tag{19b}$$

in place of (2b). Using (19a,b), our finite-difference approximation to the Poisson equation $u_{xx} + u_{yy} = f(x, y)$, at P, becomes

$$\boxed{\begin{aligned}&\frac{2}{\gamma(\gamma+\alpha)}U_W + \frac{2}{\delta(\delta+\beta)}U_S + \frac{2}{\alpha(\alpha+\gamma)}U_E\\ &\qquad\qquad + \frac{2}{\beta(\beta+\delta)}U_N - 2\frac{\alpha\gamma+\beta\delta}{\alpha\beta\gamma\delta}U_P = h^2 f_P.\end{aligned}} \tag{20}$$

Naturally, if $\alpha = \beta = \gamma = \delta = 1$, then (20) reduces to (5).

EXAMPLE 2. To illustrate the use of (20), let us solve the Poisson problem shown in Fig. 9a, using the grid shown in Fig. 9b. Obviously, the grid is quite coarse, but it should suffice for the purpose of illustration. In this case the source term is $(f(x, y) = 5x - y$ and we have chosen $h = 2$. Since there are not many grid points it will be more convenient to denote the points as $a, b, \ldots, r, s$ rather than with the double subscript notation. We need to write (20) for each of the internal grid points o, p, q, r, s. Thus, we need to determine $\alpha, \beta, \gamma, \delta$ for each of these points. At r, for instance, we compute α from $\overline{rg} \equiv \alpha h$ (where $\overline{rg}$ denotes the length of the line rg). Thus, $\alpha = \overline{rg}/h = (x_g - x_r)/h = (5.5-4)/2 = 0.75$. Next, $\overline{rg} \equiv \beta h$ gives $\beta = \overline{rg}/h = h/h = 1$, $\overline{rk} \equiv \gamma h$ gives $\gamma = (x_r - x_k)/h = (4 - \sqrt{4^2 - 2^2})/2 \approx 0.27$, and $\overline{rj} \equiv \delta h$ gives $\delta = 1$. Similarly at o, p, and s, so we have these values:

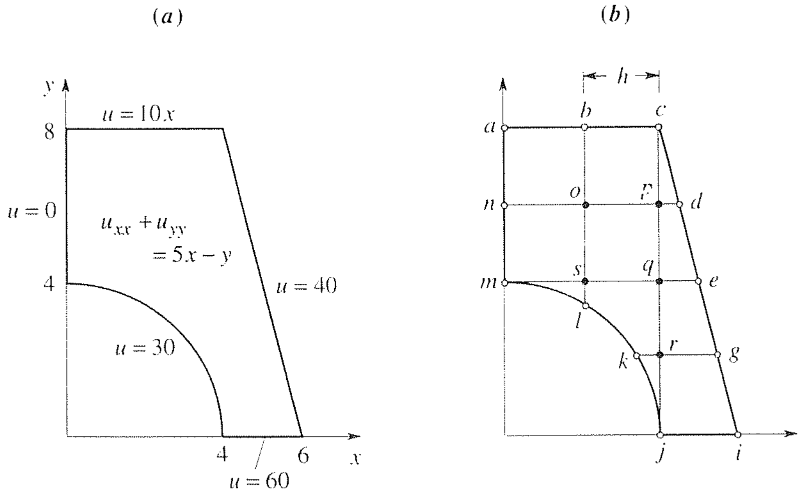


Figure 9. Example 2.

$$\begin{aligned}
o:&\quad \alpha = \beta = \gamma = \delta = 1,\\
p:&\quad \alpha = 0.25,\ \ \beta = \gamma = \delta = 1,\\
q:&\quad \alpha = 0.5,\ \ \beta = \gamma = \delta = 1,\\
r:&\quad \alpha = 0.75,\ \ \beta = 1,\ \ \gamma \approx 0.27,\ \ \delta = 1,\\
s:&\quad \alpha = \beta = \gamma = 1,\ \ \delta \approx 0.27.
\end{aligned}$$

Thus, writing (20) at these points gives the equations

$$o: \quad U_n + U_s + U_p + U_b - 4U_o = 2^2[5(2) - 6], \tag{21a}$$

$$p: \quad \frac{2}{1.25}U_o + U_q + \frac{2}{0.25(1.25)}U_d + U_c - 2\frac{1.25}{0.25}U_p = 2^2[5(4) - 6], \tag{21b}$$

$$q: \quad \frac{2}{1.5}U_s + U_r + \frac{2}{0.5(1.5)}U_e + U_p - 2\frac{1.5}{0.5}U_q = 2^2[5(4) - 4], \tag{21c}$$

$$r: \quad \frac{2}{0.27(1.02)}U_k + U_j + \frac{2}{0.75(1.02)}U_g + U_q - 2\frac{1.20}{0.20}U_r = 2^2[5(4) - 2], \tag{21d}$$

$$s: \quad U_m + \frac{2}{0.27(1.27)}U_l + U_q + \frac{2}{1.27}U_o - 2\frac{1.27}{0.27}U_s = 2^2[5(2) - 4]. \tag{21e}$$

With $U_n = 0$, $U_b = 20$, $U_c = U_d = U_e = U_g = 40$, $U_j = (60 + 30)/2 = 45$, $U_k = U_l = 30$, and $U_k = U_l = 30$, and $U_m = (30 + 0)/2 = 15$ from the boundary conditions (where we've used average values at the corners j and m, as suggested in Section 18.6), (21) becomes

$$-4U_o + U_p + U_s = -4, \tag{22a}$$

$$1.6U_o - 10U_p + U_q = -240, \tag{22b}$$

$$U_p - 6U_q + U_r + 1.33U_s = -42.7, \tag{22c}$$

$$U_q - 12U_r = -295.5, \tag{22d}$$

$$1.57U_o + U_q - 9.41U_s = -166, \tag{22e}$$

with the solution

$$U_o = 13.6,\ \ U_p = 28.3,\ \ U_q = 21.1,\ \ U_r = 26.4,\ \ U_s = 22.2. \tag{23}$$

COMMENT. If these values don't look correct, relative to the given boundary values, don't forget that besides the boundary condition inputs there is also the internal source distribution $f(x, y) = 5x - y$. Mathematically, we call the forcing term f in $\nabla^2 u = f$ a "source" term. However, if we think of this problem in terms of steady-state heat conduction, then we need to recall, from (39) in Section 16.8, that the heat source term there has a minus sign in front of it. Therefore, the $f(x, y) = 5x - y$ term in our PDE is, in physical terms, a heat *sink*, and it is the presence of that heat sink distribution that causes the interior temperatures to be lower than we would expect if we considered only the boundary conditions. For instance, if there were no sink distribution ($f = 0$) then we would expect U_p to be around 35, rather than the value 28.3 given in (23). ▮

20.5.3. Iterative algorithms. (Optional) The resulting systems of linear algebraic equations on the U_{jk}'s are of the form $\mathbf{AU} = \mathbf{c}$, where $\mathbf{A}$, typically, is quite large. To illustrate, suppose that the domain is square (as in Example 1) and that $j = 0, 1, 2, \dots, N$ and $k = 0, 1, 2, \dots, N$. Then there are $(N-1)^2$ unknown U_{jk}'s and $\mathbf{A}$ is $(N-1)^2 \times (N-1)^2$. For instance, in Example 1, $N = 4$ so $\mathbf{A}$ is 9×9. If, for greater accuracy, we choose $N = 25$, say, then $\mathbf{A}$ is 576×576. Thus, it is important to develop *efficient* methods of solution of the typically large systems of linear algebraic equations that arise.

Recall that a similar difficulty arose in Section 18.6, where the choice of an implicit (rather than explicit) scheme led to coupled (tridiagonal) systems of linear algebraic equations. However, that situation was not nearly as difficult since the problem was of initial-value type and we merely needed to solve for one time-line of unknowns at a time. Thus, with $N = 25$, say, the $\mathbf{A}$ matrix was only 24×24 rather than 576×576! It's true that we need to solve these 24th-order systems for each time step, but (supposing that there are 24 time steps, say) it is much easier to solve twenty four 24th-order systems than to solve one 576th-order system.*

Fortunately, the $\mathbf{A}$ matrix is strongly diagonal, so that we can use the same iterative techniques that were described in Section 18.6. For instance, suppose that the rectangular grid fits the domain so that we can use the scheme (5) rather than its generalization (20). Because the -4 coefficient of U_{jk} dominates the other coefficients on the left-hand side of (5) we can, to a first approximation, write $-4U_{jk} \approx h^2 f_{jk}$. Solving the latter for the U_{jk}'s and calling them $U_{jk}^{(0)}$ gives

$$U_{jk}^{(0)} = -\frac{h^2}{4} f_{jk}. \tag{24}$$

Next, we put those values into the thus-far-neglected first four terms on the left-hand side of (5), transpose them to the right, and obtain the improved values

$$U_{jk}^{(1)} = \frac{1}{4}\left(U_{j-1,k}^{(0)} + U_{j,k-1}^{(0)} + U_{j+1,k}^{(0)} + U_{j,k+1}^{(0)} - h^2 f_{jk}\right).$$

Repeating this process gives the **Jacobi** iterative scheme

$$\boxed{U_{jk}^{(n+1)} = \frac{1}{4}\left(U_{j-1,k}^{(n)} + U_{j,k-1}^{(n)} + U_{j+1,k}^{(n)} + U_{j,k+1}^{(n)} - h^2 f_{jk}\right)} \tag{25}$$

*If the truth behind this claim is not obvious to you, try changing the numbers: wouldn't you rather solve twenty 2nd-order systems than one 40th-order system?

for $n = 0, 1, 2, \dots,$ with the "starting values" given by (24).[†] As discussed in Section 18.6 for the diffusion version of (25), we can improve on (25) by using the latest iterates as soon as they become available and moving systematically across the first row (left to right), then the second, and so on. Known as the **Gauss–Seidel** or **Liebmann** method, it is expressed as

$$\boxed{U_{jk}^{(n+1)} = \frac{1}{4}\left(U_{j-1,k}^{(n+1)} + U_{j,k-1}^{(n+1)} + U_{j+1,k}^{(n)} + U_{j,k+1}^{(n)} - h^2 f_{jk}\right).} \tag{26}$$

The latter converges more rapidly than the Jacobi method and is more readily programmed.

Re-expressing (26) as

$$\begin{aligned} U_{jk}^{(n+1)} &= U_{jk}^{(n)} + \frac{1}{4}\left(U_{j-1,k}^{(n+1)} + U_{j,k-1}^{(n+1)} + U_{j+1,k}^{(n)} + U_{j,k+1}^{(n)} + U_{jk}^{(n)} - h^2 f_{jk}\right) \\ &\equiv U_{jk}^{(n)} + \Delta U_{jk}^{(n)}, \end{aligned} \tag{27}$$

we can insert a numerical control parameter ω as follows,

$$\boxed{U_{jk}^{(n+1)} = U_{jk}^{(n)} + \omega \Delta U_{jk}^{(n)},} \tag{28}$$

and choose ω so as to speed the convergence. It has been shown that the optimal value* of ω is

$$\omega_{\text{opt}} = 2\frac{1 - \sin(\pi/N)}{\cos^2(\pi/N)}. \tag{29}$$

For large N, $\omega_{\text{opt}} \sim 2$. Since $\omega > 1$ amounts to an overcorrection, in (28), the method (28) is called **successive overrelaxation**, or SOR, for brevity. We will omit a numerical illustration of these methods because the ideas are the same as for the diffusion equation; see Example 3 of Section 18.6.

One final point. We stated that the $\mathbf{A}$ matrix is strongly diagonal, but that situation will be obtained only if we write the scalar equations on the U_{jk}'s in the correct sequence, which we clarify by means of the following example.

EXAMPLE 3. Applying (5) to the Laplace problem shown in Fig. 10 gives the scalar equations

$$\begin{aligned} a:&\quad -4U_a + U_b = -100, && (30a)\\ b:&\quad U_a - 4U_b + U_c = -50, && (30b)\\ c:&\quad U_b - 4U_c + U_d = -30, && (30c)\\ d:&\quad U_c - 4U_d = -90 && (30d) \end{aligned}$$

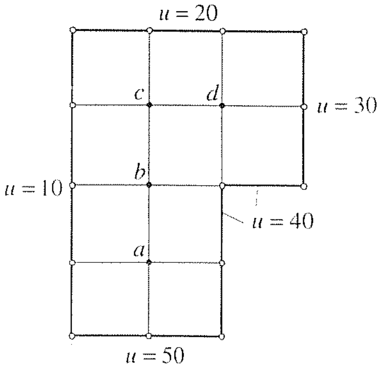


Figure 10. Example 3.

[†] Or, equivalently and more easily programmed, we can take $U_{j,k}^{(0)} = 0$. Then (25) gives $U_{j,k}^{(1)} = -(h^2/4)f_{jk}$, which is identical to the right-hand side of (24), so the subsequent iterates are the same as before.

*S. P. Frankel, "Convergence Rates of Iterative Treatments of Partial Differential Equations," *Mathematical Tables and other Aids to Computation*, Vol. 4, 1950, pp. 65–75.

or, in matrix form,

$$\begin{bmatrix} -4 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} U_a \\ U_b \\ U_c \\ U_d \end{bmatrix} = \begin{bmatrix} -100 \\ -50 \\ -30 \\ -90 \end{bmatrix}. \tag{31}$$

However, if we interchange the scalar equations (30a) and (30c), say, then we obtain, in place of (31), the equivalent system

$$\begin{bmatrix} 0 & 1 & -4 & 1 \\ 1 & -4 & 1 & 0 \\ -4 & 1 & 0 & 0 \\ 0 & 0 & 1 & -4 \end{bmatrix} \begin{bmatrix} U_a \\ U_b \\ U_c \\ U_d \end{bmatrix} = \begin{bmatrix} -30 \\ -50 \\ -100 \\ -90 \end{bmatrix}. \tag{32}$$

Thus, whereas the matrix in (31) is strongly diagonal, the one in (32) is not.

To obtain the strongly diagonal form, in any given example, proceed as follows. Order the elements of the $\mathbf{U}$ vector (in $\mathbf{AU} = \mathbf{c}$) any way you like, but then be sure to write the scalar equations in the same order, as we did in (30) and (31). ▮

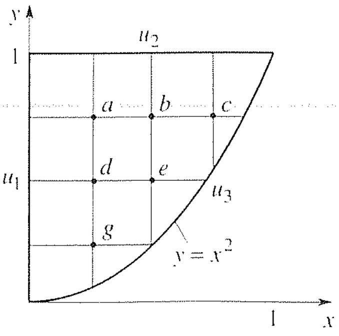
Closure. We derive the finite-difference schemes (4) and (if $\Delta x = \Delta y = h$) (5), and the generalization (20) for curvilinear boundaries. We illustrate their implementation in Examples 1 and 2, respectively, using coarse grids for simplicity. Efficient iterative solution techniques, that are needed for fine grids (i.e., for large $\mathbf{A}$ matrices), are described in the optional Section 20.5.3. Dirichlet boundary conditions are the simplest and are used throughout, although other boundary conditions are considered in the exercises.

EXERCISES 20.5

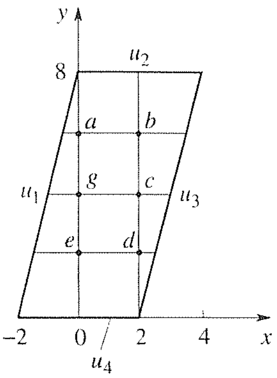
1. Consider the Poisson problems $u_{xx}+u_{yy}=f(x,y)$ labeled A, B, C, D in the accompanying figures, each with Dirichlet boundary conditions. Write the linear algebraic equations governing the unknown nodal values and solve for those values, either by hand or by computer.

(a) Problem A with $u_1(y)=10y,\ u_2=u_3=20,\ f(x,y)=-20xy$.
(b) Problem A with $u_1=u_2=0,\ u_3=50,\ f(x,y)=0$.
(c) Problem A with $u_1=u_2=u_3=0,\ f(x,y)=-100$.
(d) Problem B with $u_1=u_2=u_3=u_4=0,\ f(x,y)=-50$.
(e) Problem B with $u_2=100,\ u_1=u_3=u_4=f(x,y)=0$.
(f) Problem B with $u_1=u_2=u_3=0,\ u_4=200,\ f(x,y)=x^2+y^2$.
(g) Problem C with $u_1=u_2=u_3=u_4=0,\ u_5=50,\ f(x,y)=-20$.
(h) Problem C with $u_1=u_2=u_3=u_4=u_5=100,\ f(x,y)=30$.
(i) Problem C with $u_1=u_3=u_4=u_5=0,\ u_2(x)=10x,\ f(x,y)=50(x^2-y^2)$.
(j) Problem D with $u_1=u_3=f(x,y)=0,\ u_2=u_4=u_5=100$. HINT: Note the symmetry about $y=4$.
(k) Problem D with $u_1=u_3=u_4=f(x,y)=0,\ u_2=50,\ u_5=-50$. HINT: Note the antisymmetry about $y=4$, so $u(x,4)=0$. You may wish to work the problem twice: first, using the noted antisymmetry and then again, this time not using it (in case you have any doubts).
(l) Problem D with $u_1=u_3=u_4=0,\ u_2=u_5=50,\ f(x,y)=10(x^2+y^2)$.

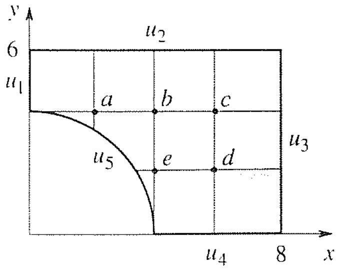
A.



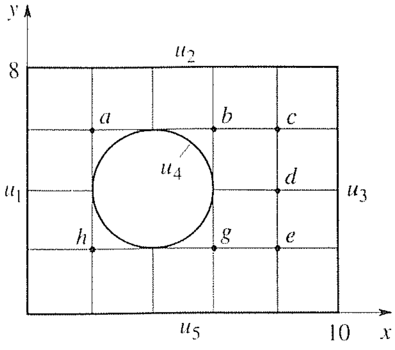
B.



C.



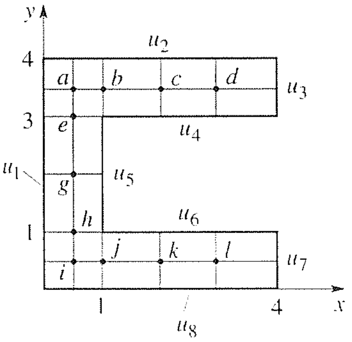
D.



2. Consider the Poisson equation $u_{xx}+u_{yy}=f(x,y)$ on the domain between two nested squares, one with corners at $(1,1),(-1,1),(-1,-1),(1,-1)$, and the other with corners at $(0.6,0.6),(-0.6,0.6),(-0.6,-0.6),(0.6,-0.6)$. Let $\Delta x=\Delta y=h=0.2$. Given $f(x,y)$ and the boundary conditions, use (5) to solve for u at each of the 32 nodal points using any symmetries or antisymmetries that are present to reduce the number of unknowns.

(a) $f(x,y) = -100$, $u = 0$ on inner and outer boundaries
(b) $f(x,y) = 0$, $u = 0$ on inner boundary, $u = 100$ on outer boundary
(c) $f(x,y) = 0, u(0.6,y) = u(1,y) = u(x,0.6) = u(x,1) = 50, u(-0.6,y) = u(-1,y) = u(x,-0.6) = u(x,-1) = -50$
(d) $f(x,y) = 0, u(-1,y) = u(-0.6,y) = u(0.6,y) = u(1,y) = 0, u(x,-1) = u(x,-0.6) = -20, u(x,0.6) = u(x,1) = 20$

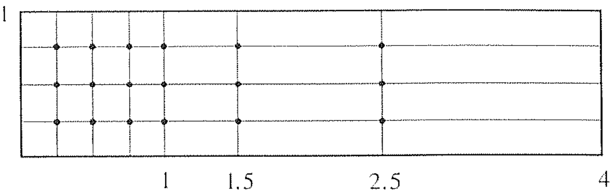
3. We developed equation (20), based on the pattern in Fig. 8, to enable us to handle the domains with irregular boundaries. However, we can also use (20) [or, equivalently, (4) with $\Delta x \neq \Delta y$] to increase the nodal point density in regions where greater resolution is needed, as indicated in the figure below. Solve for $U_a, U_b, \ldots, U_l$ using any symmetries or antisymmetries that are present to reduce the number of unknowns. HINT: At a use $h = 0.5$, at b use $h = 1$ and so on.



(a) $f(x,y) = -10$, $u_1 = \cdots = u_8 = 0$
(b) $f(x,y) = 0$, $u_1 = u_2 = u_4 = u_5 = u_6 = u_8 = 0$, $u_3 = u_7 = 50$
(c) Same as (b), but with u_7 changed to -50.

4. First, read Exercise 3. For the Laplace equation on the domain shown, with the Dirichlet boundary conditions $u(0,y) = 100$, $u(x,0) = u(x,1) = u(4,y) = 0$, $u(x,y)$ will vary rapidly only near the left end. Thus, let us bunch the nodal points as shown.

(a) Solve for u at the 21 nodal points using the finite-difference method. HINT: Use symmetry to reduce the number of unknowns to 12.
(b) Compare your computed values at the six nodal points on the horizontal centerline $y = 0.5$ with the exact values, obtained by separation of variables.



5. Show that (20) agrees with (4), as it should, if we set $\alpha h = \gamma h = \Delta x$ and $\beta h = \delta h = \Delta y$.

6. (*Empirical estimate of the order of the method*) (a) Show that the truncation error in (5) is $O(h^2)$ so the method is of second order.
(b) To test the assertion in part (a), note that if the method is of order p that means that

$$u(x_j, y_k) - U_{j,k} \sim Ch^p \tag{6.1}$$

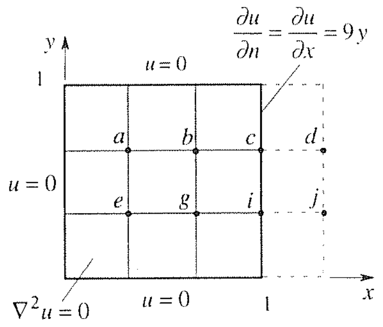
as $h \to 0$, where x_j, y_k is any fixed field point within the domain, $u(x_j, y_k)$ is the exact solution there, $U_{j,k}$ is the computed solution there, and C is some constant. It suffices to use a concrete example. For that purpose, use the point $x = y = 0.5$ in Example 1. We found that $u(0.5, 0.5) = 19.9$, with $h = 0.5$ we obtained $U_{1,1} = 25$ there, and with $h = 0.25$ we obtained $U_{2,2} = 21.3$ there. Writing (6.1) for each of these two cases gives two equations in the unknowns C and p. Solving for p, show that $p \approx 1.87$. NOTE: In (6.1) $U_{j,k}$ denotes the value obtained using a perfect computer, one with no roundoff error, and that is virtually the case since the roundoff error should be extremely small compared to the truncation error.
(c) In obtaining $p \approx 1.87$, in part (b), we treated (6.1) as an equation (i.e., with an equal sign), whereas it is only true as $h \to 0$. Thus, we can expect a better empirical estimate of p if we use two successive small values of h such as 1/4 and 1/6, rather than 1/2 and 1/4. Here, we ask you to compute U at $x = 0.5$, $y = 0.5$ using $h = 1/6$ (in which case there will be 25 equations in 25 unknowns, which can be reduced to 15 equations in 15 unknowns by using symmetry) and to use the $h = 1/4$ and $h = 1/6$ results to obtain a more accurate estimate of p. NOTE: Remember that when we say h is small we mean small relative to the size of the domain. In this case the domain is a unit square so the values $h = 1/4$ and $h = 1/6$ are not especially small. If, instead, the domain were of dimension 50 × 50, then these h values would be quite small.

7. (*Neumann and Robin boundary conditions*) First, recall the forward, backward and central difference quotients.

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}, \qquad \text{(forward)}$$

$$f'(x) \approx \frac{f(x) - f(x-h)}{h}, \qquad \text{(backward)}$$
$$f'(x) \approx \frac{f(x+h) - f(x-h)}{2h}, \qquad \text{(central)}$$
$$(7.1,2,3)$$

with truncation errors that are $O(h)$, $O(h)$, and $O(h^2)$, respectively. Thus far in this section we have considered boundary conditions only of Dirichlet type. To see how to handle a Neumann condition consider the representative problem shown in the figure.



(a) Writing out equation (5) at a, b, e, g gives four equations in six unknowns U_a, U_b, U_c, U_e, U_g, and U_i. To apply the Neumann condition $u_x(1, y) = 9y$ at c and i, use the backward difference quotient (7.2), show that the resulting system is

$$\begin{bmatrix} -4 & 1 & 0 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} U_a \\ U_b \\ U_c \\ U_e \\ U_g \\ U_i \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad (7.4)$$

and solve for $U_a, \ldots, U_i$.

(b) However, whereas the first, second, fourth, and fifth scalar equations in (7.4) [namely, those resulting from the application of (5)] have a truncation error that is $O(h^2)$ the third and sixth [those resulting from the application of (7.2) to the Neumann boundary condition] have a truncation error that is $O(h)$. Just as "the chain is as weak as the weakest link," these $O(h)$ errors contaminate the whole system (7.4) and cause an overall truncation error that is $O(h)$. To prevent this reduction in accuracy we can use the central difference quotient (7.3) instead of the backward difference quotient (7.2). The idea is to extend the domain as indicated by the dashed lines. Applying (5) at a, b, c, e, g, i gives six equations in the eight unknowns. To apply the Neumann condition at c and i use (7.3), thereby obtaining two more equations. The result is eight linear algebraic equations on $U_a, \ldots, U_j$. Obtain that system and solve it for $U_a, \ldots, U_j$. NOTE: Of course, in the end you can discard the auxiliary values U_d and U_g.

(c) Solve the problem exactly, by separation of variables, and compare your values of U_c, say, from parts (a) and (b), with the exact values at a, b, e, and g.

(d) With the Neumann boundary condition $u_x(1, y) = 9y$ changed to the Robin boundary condition

$$u_x(1, y) + 3u(1, y) = 9y,$$

modify (7.4) accordingly.

8. (*Other elliptic PDE's*) The Laplace and Poisson equations are *elliptic*, and the methods developed in this section are applicable to other elliptic PDE's as well. In each case, first verify that the PDE is elliptic. Then derive a finite-difference scheme (with $\Delta x = \Delta y = h$) analogous to (5), using (2a), (2b), and central difference quotients (see Exercise 7) for first-order derivatives. Then, apply the finite-difference scheme at each interior node for the case where the domain is $0 < x < 1$, $0 < y < 1$ with $u(x, 0) = 0$, $u(0, y) = 20$, $u(x, 1) = 50$, $u(1, y) = 10y$, and with $h = 1/3$. (You need not solve the resulting system of equations.)

(a) $(1 + x^2)u_{xx} + u_{yy} = 0$
(b) $u_{xx} + 2u_{yy} - u_x = 4$
(c) $u_{xx} + u_{yy} - u = 20x - 5y$
(d) $u_{xx} + (1 + x^2 + y^2)u_{yy} + 2u_y = 15(x^2 + y^2)$

Chapter 20 Review

Recall that the PDE

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f, \tag{1}$$

where $A, \ldots, F, f$ may be functions of x and y, is classified as hyperbolic, parabolic, or elliptic as follows:

$$\begin{array}{lll} \text{hyperbolic} & \text{if} & B^2 - AC > 0, \\ \text{parabolic} & \text{if} & B^2 - AC = 0, \\ \text{elliptic} & \text{if} & B^2 - AC < 0, \end{array} \tag{2}$$

the classification depending only on the coefficients A, B, C of the second-order derivatives. The prototypical equations considered in these chapters are the hyperbolic wave equation $c^2u_{xx} - u_{tt} = 0$ (with t in place of y), the parabolic diffusion equation $\alpha^2u_{xx} - u_t = 0$, and the elliptic Laplace equation $u_{xx} + u_{yy} = 0$.

Observe from (2) that the parabolic case is on the borderline between the hyperbolic and elliptic cases, so it is not surprising to learn, in these chapters, that parabolic equations share some properties with hyperbolic equations and some with elliptic equations. For example:

(1) *Both the hyperbolic wave equation and the parabolic diffusion equation are of initial-value type (in the t variable), whereas the elliptic Laplace equation is of boundary-value type.* For instance for a wave or diffusion problem defined on $0 < x < L$ and $0 < t < \infty$ the solution at any x, t point depends on the initial data (at $t = 0$) and on the boundary data up to time t but not beyond, whereas for a Laplace problem defined on $0 < x < L$ and $0 < y < \infty$ the solution at any x, y point depends on the data at $y = 0$ and on the data at $x = 0$ and $x = L$ over the entire interval $0 < y < \infty$. The initial-value/boundary-value distinction is especially important in numerical solution by the finite-difference method. For instance, contrast the finite-difference solutions of a diffusion problem on $0 < x < L$, $0 < t < T$ with $\Delta x = L/N$ and $\Delta t = T/M$, and a Laplace problem on $0 < x < L$, $0 < y < Y$ with $\Delta x = L/N$ and $\Delta y = Y/M$. Because of its initial-value type, the diffusion problem admits an explicit solution or, at worst, requires the solution of an $(N-1)$th-order matrix equation for each time-line of nodal values. In contrast, the boundary-value type Laplace problem requires us to solve for all the unknown nodal values at once; that is, we need to solve an $[(M-1)(N-1)]$th-order matrix equation.

(2) *For the hyperbolic wave equation discontinuities or kinks in initial or boundary data propagate into the solution domain, whereas for the parabolic diffusion equation and the elliptic Laplace equation they do not; they are smoothed upon "entering" the solution domain.* For instance, contrast Fig. 4 in Section 18.4, where the initial condition is a Heaviside step function and where the resulting solution is a smooth (even infinitely differentiable) function of

x for all $t > 0$, and Fig. 4 in Section 19.4, where the initial deflection is "kinky" and where those kinks propagate into the solution domain.

We can summarize these comparisons, for mnemonic purposes, in the following tabular form.

Wave Equation:	kinks propagate,	initial-value type
Diffusion Equation:	smooth,	initial-value type
Laplace equation:	smooth,	boundary-value type

Let us return from this comparison of the wave, diffusion, and Laplace equations to the present chapter on the Laplace equation. Included are the solutions to particularly well known Dirichlet problems, namely, for the rectangle (Section 20.2), circular disk (Section 20.3.1), circular cylinder with axisymmetry (Section 20.3.2), sphere with axisymmetry (Section 20.3.3), and half plane (Section 20.4).

Prominent theoretical results include the average value property and maximum principle. The *average value property* is that if $\nabla^2 u = 0$ in a two-dimensional (or three-dimensional) domain $\mathcal{D}$, then u at any point P' within $\mathcal{D}$ is equal to the average value of u around any circle (or on any sphere) centered at P' and lying entirely within $\mathcal{D}$. And the **maximum (minimum) principle** is that u cannot attain its maximum (or minimum) value in $\mathcal{D}$ (unless u is a constant everywhere); it must attain its maximum (or minimum) on the boundary of $\mathcal{D}$.

Chapter 21

Functions of a Complex Variable

21.1 Introduction

There is much to recommend the study of complex variable theory. In this text on applied mathematics, we should state first that the subject is of great importance in applications. For example, it may be recalled from our discussion of the Laplace transform that the inverse Laplace transformation is given as a contour integral in a complex s plane. Not yet having studied complex variable theory, we were forced to avoid the inversion formula and to rely, instead, on the use of transform tables. More generally, the evaluation of a wide class of definite integrals (even along the real axis) is facilitated by use of the complex integral calculus. Another particularly important application is in the use of conformal mapping to solve boundary-value problems in two-dimensional potential theory (i.e., governed by the two-dimensional Laplace equation).

Also important about complex variable theory is that it serves to "complete" our understanding, from the calculus, of real-valued functions of a single real variable. For example, in studying Taylor series one finds that the expansion

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots$$

holds only for $|x| < 1$. Yet the function $1/(1+x^2)$ is infinitely differentiable for *all* x, and offers no clue as to why the interval of convergence should be anything less than infinite. It is only when we consider instead the function $1/(1+z^2)$, where $z = x + iy$, that the source of the difficulty comes into view, namely, singularities at $z = \pm i$, off of the real axis, in the complex z plane.

Our sequence of topics echoes the usual format of the real variable calculus. Beginning in Chapter 21 with a discussion of *complex numbers* and the complex plane, we next introduce the notion of a complex valued *function* of a (single) complex variable and then define and discuss a number of *elementary functions*, such as the exponential, trigonometric, and hyperbolic functions. Defining a *limit concept* and *continuity*, we are able in the final section to define a particularly important limit, the *derivative*. Thus, in this chapter, we get as far as the differential calculus of functions of a complex variable.

The complex integral calculus and series expansions are then developed in Chapters 23 and 24. In between is Chapter 22 on conformal mapping. That chapter is not a prerequisite for Chapters 23 or 24. Thus, in a shorter course one might cover only Chapters 21 and 22, or, instead, Chapters 21, 23, and 24.

21.2 Complex Numbers and the Complex Plane

Historically, complex numbers were created several hundred years ago, within the context of the theory of equations. For if one allowed only real numbers, then equations such as $x^2 + 1 = 0$ and $x^2 + 2x + 4 = 0$ had no solution. Thus, in a step that was slow to gain general acceptance, a broader number system was devised so that the equations given above, and indeed every polynomial equation, possess solutions within that number system.* Eventually named **complex numbers** by *Carl Friedrich Gauss* (1777–1855), these new numbers were of the form $a + ib$ where a and b are real and where i satisfies the equation $i^2 = -1$. It is important to understand that the plus sign in $a + ib$ does not denote addition; rather, $a + ib$ is a single number, not the sum of a and ib.

The (real) numbers a, b are called the **real part** and **imaginary part** of $a + ib$, respectively:

$$\text{Re}(a + ib) \equiv a, \qquad \text{Im}(a + ib) \equiv b \quad (\text{not } ib). \tag{1}$$

We do not distinguish between $a + ib$ and $a + bi$, and we generally write $a + i0$ as a, and $0 + ib$ as ib, for brevity. The former complex number is said to be *purely real* and the latter is said to be *purely imaginary*. Finally, two complex numbers are said to be *equal* if their real and imaginary parts, respectively, are equal: that is,

$$a_1 + ib_1 = a_2 + ib_2 \tag{2}$$

holds if and only if $a_1 = a_2$ and $b_1 = b_2$.

Beyond introducing complex numbers, we will need an algebra for their manipulation. The idea will be to stay as close as possible to the rules of ordinary arithmetic (i.e., governing real numbers). For example, the rules of ordinary arithmetic would seem to dictate that

$$(a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2)$$

and that

$$\begin{aligned}(a_1 + ib_1)(a_2 + ib_2) &= a_1a_2 + ia_1b_2 + ib_1a_2 + i^2b_1b_2 \\ &= (a_1a_2 - b_1b_2) + i(a_1b_2 + b_1a_2).\end{aligned}$$

*Perhaps the reluctance to accept complex numbers can be better appreciated if we mention that, before complex numbers, there had even been reluctance to accept *negative* numbers. After all, how could one have a negative amount of something. (Of course, that was before the invention of the credit card.)

But, be sure to see that these equalities have no logical support since, as noted above, $a_1 + ib_1$ is a single complex number, not a_1 plus ib_1; similarly for $a_2 + ib_2$. Rather, we have merely been trying to motivate reasonable definitions for the addition and multiplication of complex numbers. Accordingly, we now *define*

$$(a_1 + ib_1) + (a_2 + ib_2) \equiv (a_1 + a_2) + i(b_1 + b_2) \tag{3}$$

and

$$(a_1 + ib_1)(a_2 + ib_2) = (a_1a_2 - b_1b_2) + i(a_1b_2 + b_1a_2). \tag{4}$$

For instance, $(3 - 2i) + (5 + i) = 8 - i$ and $(3 - 2i)(5 + i) = 17 - 7i$.

Before proceeding, let us agree to denote complex numbers by a single letter, usually z, for brevity. Thus, we write $z = a + ib$. With the definitions above, it is readily verified that the familiar rules of algebra hold for complex numbers. For instance, if we denote any three complex numbers as $z_1 = a_1 + ib_1$, $z_2 = a_2 + ib_2$, and $z_3 = a_3 + ib_3$, then

$$\begin{aligned} z_1 + z_2 &= z_2 + z_1, & z_1z_2 &= z_2z_1 & &\text{(commutative)} & (5)\\ (z_1 + z_2) + z_3 &= z_1 + (z_2 + z_3), & (z_1z_2)z_3 &= z_1(z_2z_3) & &\text{(associative)} & (6)\\ z_1(z_2 + z_3) &= z_1z_2 + z_1z_3 & & & &\text{(distributive).} & (7) \end{aligned}$$

Further, there are zero and unit complex numbers since, from (3),

$$z + (0 + i0) = z$$

and, from (4),

$$(1 + i0)z = z$$

for all z. Thus, for brevity, we write $0 + i0 = 0$ and $1 + i0 = 1$. Finally, $-z \equiv (-1)z$, and the *subtraction* of complex numbers is defined, in terms of the already-defined operations of addition and multiplication, by $z_1 - z_2 \equiv z_1 + (-z_2) = z_1 + (-1)z_2$.

It is to be stressed that mathematical notation is important, and that the foregoing discussion can be reorganized by introducing complex numbers as **ordered pairs** of real numbers. Thus, in place of $z = a + ib$ we could write

$$z = (a, b),$$

where a and b are called the real and imaginary parts, respectively, of the complex number z. If $z_1 = (a_1, b_1)$ and $z_2 = (a_2, b_2)$ are any two complex numbers, we define their sum and product as

$$z_1 + z_2 = (a_1, b_1) + (a_2, b_2) \equiv (a_1 + a_2, b_1 + b_2) \tag{8}$$

and

$$z_1z_2 = (a_1, b_1)(a_2, b_2) \equiv (a_1a_2 - b_1b_2, a_1b_2 + a_2b_1), \tag{9}$$

respectively. These definitions are equivalent to (3) and (4). Although the form $z = a + ib$ is more convenient, and will be used almost exclusively in the sequel, the ordered-pair approach, created by the great Irish mathematician *William*

R. Hamilton (1805–1865), is important. First, it is based entirely on real numbers and completely avoids the mysterious "imaginary" number $i = \sqrt{-1}$. Since the forms $z = a + ib$ and $z = (a, b)$ are found to be equivalent, we can thus put to rest apprehensions about the quantity i and accept it as a computational convenience. Second, the ordered-pair format $z = (a, b)$ suggests a graphical representation of z as a point in a Cartesian a, b plane. Since Cartesian axes are more usually denoted by x and y rather than by a and b, let us write $z = x + iy = (x, y)$ and represent z as a point in a so-called **complex z plane**, as shown in Fig. 1.* The x axis is called the **real axis**, and the y axis is called the **imaginary axis**.†

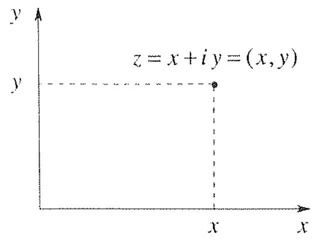


Figure 1. Complex z plane.

Observe from Fig. 2 that the addition of complex numbers defined by (3) [or, equivalently, by (8)] satisfies the parallelogram law for the addition of vectors so that it is often convenient to think of complex numbers as *vectors*. That is, a z vector is the vector from the origin to the point z.

The distance from the origin to the point z (i.e., the "length of the z vector") is called the **modulus** of z and, by analogy with the absolute value of a real number, is denoted as $|z|$, or as $\text{mod}(z)$. Clearly (Fig. 1),

$$|z| = \sqrt{x^2 + y^2}. \tag{10}$$

For example, $|2 - i| = \sqrt{5}$. It is easy to verify (Exercise 2) that

$$|z_1 z_2| = |z_1|\,|z_2|\,; \tag{11}$$

that is, the modulus of a product equals the product of the moduli of the factors. Further (Fig. 3), the inequality

$$|z_1 + z_2| \leq |z_1| + |z_2| \tag{12}$$

follows from the Euclidean proposition that the length of any one side of a triangle is less than or equal to the sum of the lengths of the other two sides. Hence (12) is known as the **triangle inequality**.

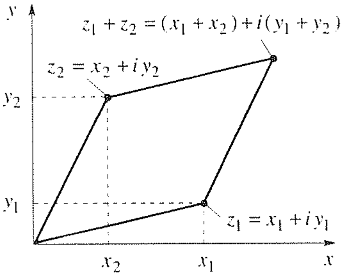


Figure 2. Addition in the z plane.

Note carefully that the complex numbers are not ordered as real numbers are. For example, whereas $-6 < 2$, $10 > 7$, and so on, analogous statements such as $z < 0$, $z > 3$, and $4 + 3i > 1 + 2i$ are not meaningful! Of course, statements such as $\text{Re}\, z < 6$, $\text{Im}\, z < 6$, $|z| > 4$, and $|1 - i| < |3 - i|$ *do* make sense because $\text{Re}\, z$, $\text{Im}\, z$, and $|z|$ are real numbers.

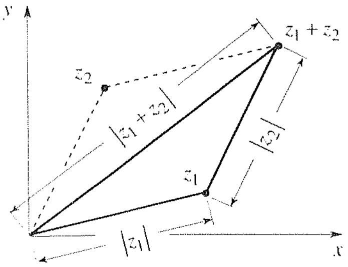


Figure 3. Triangle inequality.

Besides $z = x + iy$, it is useful to define the **complex conjugate** of z as

$$\bar{z} \equiv x - iy. \tag{13}$$

*The complex plane is sometimes called the *Argand diagram* after the French mathematician *Jean Robert Argand* (1768–1822), who was one of the first to propose such representation of complex numbers. A bookkeeper by profession, Argand was a self-taught mathematician.

†One might be concerned that we have labeled the point on the y axis as "y." Shouldn't it be "iy?" The answer is that either the real or the complex label is correct, depending on whether we are regarding the point as a point on the real y axis or as a point in the complex z plane.

Thus, if $z_1 = 8 + 3i$ and $z_2 = -4i$, then $\bar{z}_1 = 8 - 3i$ and $\bar{z}_2 = 4i$. It is readily shown (Exercise 3) that

$$\overline{z_1 + z_2} = \bar{z}_1 + \bar{z}_2, \tag{14a}$$

$$\overline{z_1 z_2} = \bar{z}_1 \bar{z}_2, \tag{14b}$$

and that

$$|z| = \sqrt{z\bar{z}}. \tag{15}$$

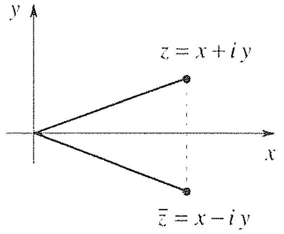


Figure 4. Complex conjugate.

Graphically, $\bar{z}$ is simply the reflection of z about the real axis (Fig. 4).

Finally, the complex conjugate is useful in defining the *division* of two complex numbers. Division is defined as the inverse of multiplication. That is, the quotient $z = z_1/z_2$ is the complex number $z = x + iy$ such that $z_2 z = z_1$. Writing out the latter as

$$(x_2 + iy_2)(x + iy) = x_1 + iy_1, \tag{16}$$

expanding the left-hand side, and equating real and imaginary parts on the left- and right-hand sides, we obtain

$$x = \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2}, \qquad y = \frac{x_2 y_1 - x_1 y_2}{x_2^2 + y_2^2}. \tag{17}$$

Practically speaking, however, it is simpler to multiply numerator and denominator by the complex conjugate of the denominator,

$$z = \frac{z_1}{z_2} = \frac{z_1}{z_2}\frac{\bar{z}_2}{\bar{z}_2} \tag{18}$$

because the denominator $z_2\bar{z}_2 = |z_2|^2$ is now real. Thus, writing out (18), we find that

$$\begin{aligned} z = \frac{z_1}{z_2} = \frac{z_1\bar{z}_2}{z_2\bar{z}_2} &= \frac{(x_1 + iy_1)(x_2 - iy_2)}{x_2^2 + y_2^2} \\ &= \frac{x_1 x_2 + y_1 y_2}{x_2^2 + y_2^2} + i\frac{x_2 y_1 - x_1 y_2}{x_2^2 + y_2^2}, \end{aligned} \tag{19}$$

which result agrees with (17).

EXAMPLE 1.

$$\frac{2+i}{3-4i} = \frac{2+i}{3-4i}\,\frac{3+4i}{3+4i} = \frac{(6-4)+(8+3)i}{9+16} = \frac{2}{25} + \frac{11}{25}i,$$

which result can be checked by showing that $3 - 4i$ times $\frac{2}{25} + \frac{11}{25}i$ gives $2 + i$. ▮

Closure. In closing this section, we note that besides the *Cartesian* notation $z = x + iy$, for complex numbers, there is also a *polar* notation "$z = re^{i\theta}$" that is often convenient. However, because it contains the complex exponential $e^{i\theta}$, we delay our introduction of the polar notation until we discuss the complex exponential function.

EXERCISES 21.2

1. Using the definitions (3) and (4), verify

(a) the commutative properties (5)
(b) the associative properties (6)
(c) the distributive property (7)

2. Verify (11).

3. Verify (14a), (14b), and (15).

4. Recall that the definitions of addition and multiplication, (3) and (4), were rigged so that the familiar commutative, associative, and distributive rules hold. Thus, one can manipulate complex numbers easily and with little fuss. Nevertheless, it may be valuable to work an example or two, paying careful attention to the definitions and properties (3) to (7). In this spirit we ask you to solve the following equations for z and then to verify, carefully, that the root(s) do indeed satisfy the given equation. In each step of the verification, identify the definition or property [(3) to (7)] being used. Naturally, z^2 means zz.

(a) $z^2 - 2z + 2 = 0$ (b) $2z^2 + 2z + 1 = 0$
(c) $3z^2 + 2z + 1 = 0$ (d) $z^2 + 2iz + 1 = 0$
(e) $10z^2 - 6z + 1 = 0$ (f) $z^2 - 3iz - 2 = 0$

5. Show that

(a) $|z^3| = |z|^3$
(b) $|z^n| = |z|^n$ and $|1/z^n| = 1/|z|^n$ for $n = 1, 2, \dots$
(c) $|z_1 z_2 z_3| = |z_1|\,|z_2|\,|z_3|$
(d) $|z_1 z_2 \cdots z_n| = |z_1|\,|z_2| \cdots |z_n|$ for $n = 1, 2, \dots$
(e) $|z_1 + z_2 + z_3| \leq |z_1| + |z_2| + |z_3|$
(f) $|z_1 + z_2 + \cdots + z_n| \leq |z_1| + |z_2| + \cdots + |z_n|$ for $n = 1, 2, \dots$

6. Show that

(a) $\overline{\overline{z}} = z$
(b) $\overline{\left(\frac{1}{z}\right)} = \frac{1}{\overline{z}}$
(c) z is real if and only if $z = \overline{z}$
(d) $\overline{z^3} = \overline{z}^3$
(e) $\overline{z^n} = \overline{z}^n$ for $n = 1, 2, \dots$

7. Show that if $z = x + iy$, then $x = (z + \overline{z})/2$ and $y = (z - \overline{z})/2i$.

8. Show that if $z_1 z_2 = 0$, then at least one of the two factors must be zero.

9. Evaluate each of the following. That is, express each in standard Cartesian form $x + iy$.

(a) $(2 - i)^3$ (b) $\dfrac{1}{1 - 2i}$
(c) $\dfrac{i}{2 + 5i}$ (d) $\dfrac{1 + i}{1 - i}$
(e) $\left(\dfrac{1 + i}{2 - i}\right)^3$ (f) $\operatorname{Re} \dfrac{2 + 3i}{4 + 5i}$
(g) $\operatorname{Im} (1 + i)^3$ (h) $\left(\operatorname{Re} \dfrac{1}{1 + i}\right)^3$
(i) $\operatorname{Im} \dfrac{a + ib}{c + id}$ (j) $\operatorname{Re} \dfrac{1}{1 - i}$

10. Evaluate

(a) $\left|\dfrac{1 - i}{1 + i}\right|$ (b) $\left|\dfrac{(2 - i)^3}{(1 + 3i)^2}\right|$
(c) $|(1 - 2i)^2 + (1 + i)^2|$

11. Verify the triangle inequality (12) for each of the following cases by working out the left- and right-hand sides.

(a) $z_1 = 2 + 3i$, $z_2 = 4 - i$
(b) $z_1 = 1 + i$, $z_2 = 7i$
(c) $z_1 = 5$, $z_2 = 4i$
(d) $z_1 = 3 + 4i$, $z_2 = 2 + i$
(e) $z_1 = 1 + i$, $z_2 = 1 - i$

21.3 Elementary Functions

21.3.1. Preliminary ideas. Having introduced complex numbers $z = x + iy$, we next introduce *functions* of a complex variable. Since functions will be defined on sets of complex numbers, we need to distinguish various topological features of these sets.

First, we define the **distance** between any two points $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ as

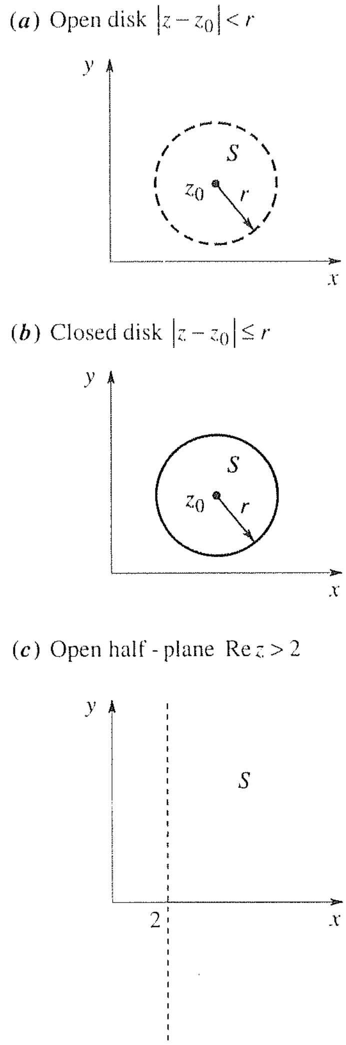
$$d(z_1, z_2) \equiv |z_1 - z_2| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{1}$$

Next, we define a **neighborhood** $N(z_0; r)$ of a point z_0 in the z plane as the set of points z closer to z_0 than r, that is, all z's such that

$$|z - z_0| < r \qquad (r > 0). \tag{2}$$

A set S is **connected** if each pair of points in S can be joined by an unbroken line consisting of a finite number of straight segments, each contained entirely within S. A point P is a **boundary point** of S if every neighborhood of P (i.e., no matter how small) contains points in S as well as points not in S, and a point P in S is an **interior point** of S if there exists some neighborhood of P lying entirely within S. Finally, a connected set S is a **region**; S is an **open region** or **domain** if it contains *none* of its boundary points, and is a **closed region** if it contains *all* of its boundary points.

To illustrate, $|z - z_0| < r$ defines an open region (Fig. 1a), $|z - z_0| \leq r$ defines a closed region (Fig. 1b), and Re $z > 2$ defines an open region (Fig. 1c).


(*a*) Open disk $|z - z_0| < r$

(*b*) Closed disk $|z - z_0| \leq r$

(*c*) Open half-plane Re $z > 2$

Figure 1. The regions $|z - z_0| < r$, $|z - z_0| \leq r$, Re $z > 2$.

Turning to functions, recall from the elementary calculus that a function f of a real variable x is a rule that assigns a unique value $f(x)$ to each point x in some set on an x-axis, as illustrated in Fig. 2. One also calls f a **mapping**. Following *Descartes* (1596–1650), a useful graphical display of f (the "graph" of f) can be obtained by arranging the x and f axes at right angles to each other and plotting the set of points $x, f(x)$, as illustrated in Fig. 3. Analogously, we define a **function w of a complex variable z** as a rule that assigns a unique value $w(z)$ to each point z in some set D in the complex plane.*

In general, $w(z)$ is complex, so the mapping is as depicted in Fig. 4 from the z plane to a w plane. The set D on which w is defined is called the **domain** of definition† of w, and the set R of $w(z)$ values is called the **range** of values of w. Strictly speaking, one distinguishes between the function (or mapping) and the *values* of the function by using the notation w for the former and $w(z)$ for the latter. However, we plead guilty of occasionally writing "the function

*Just as the letter f is traditionally, although by no means exclusively, used for functions of a real variable, it is traditional to use either w or f for functions of a complex variable. We plan to use both extensively, though not exclusively.

†The domain of definition of a function is not necessarily an open set and is thus not necessarily a domain in the sense defined above.

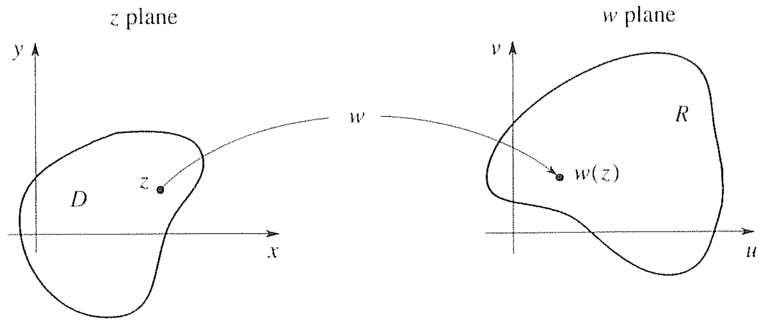


Figure 4. The mapping w.

$w(z)$" as a shorthand way of writing "the function w whose values are $w(z)$."

Notice carefully that Descartes's method of graphical display (Fig. 3) is not available to us in the complex case because both D and R are two-dimensional so that the "graph of w" would require a plot in four dimensions. *Thus, graphs like the one shown in Fig. 3 will not be possible for a function of a complex variable.* Of course, the real and imaginary parts of $w(z)$, say u and v, respectively, must be functions of x and y. Thus, we can express

$$w(z) = u(x, y) + i\,v(x, y), \tag{3}$$

and although we cannot plot "the graph of w," we can do three-dimensional plots of u and v as functions of x and y, but this form of display is seldom used.

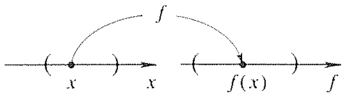


Figure 2. The mapping f.

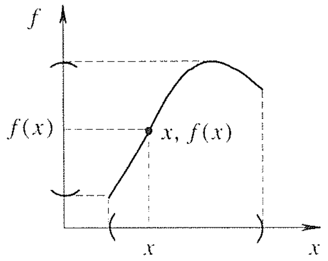


Figure 3. The graph of f.

EXAMPLE 1. To illustrate some of these ideas, consider $w(z) = z^2$, defined on the first quadrant of the z plane: $0 < x < \infty$, $0 < y < \infty$. Then $w(z) = (x + iy)^2 = x^2 - y^2 + i2xy$ so that $u(x, y) = x^2 - y^2$ and $v(x, y) = 2xy$. Since $0 < x < \infty$ and $0 < y < \infty$, it follows from the form of u and v that $-\infty < u < \infty$ and $0 < v < \infty$ so that the range of w is the entire upper half plane $v > 0$ as shown in Fig. 5.

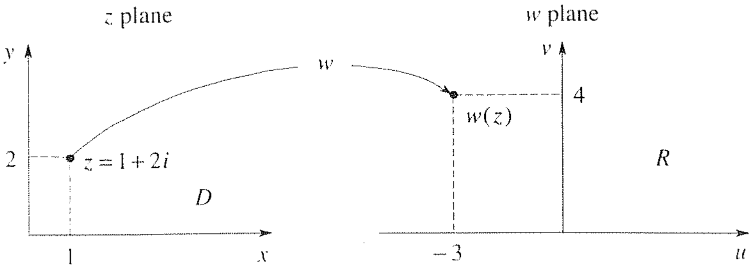


Figure 5. Mapping defined in Example 1.

If, for example, $z = 1 + 2i$, then $w(z) = (1 - 4) + i2(1)(2) = -3 + 4i$ as shown in the figure.

Since we cannot draw a graph of $w(z) = z^2$ [as we could for the real function $w(x) = x^2$], and we do not wish to plot the surfaces $u(x, y) = x^2 - y^2$, $v(x, y) = 2xy$

(although this could be done using computer graphics), the question remains: what *can* be done, easily, to provide some form of graphical display of w? We could, of course, select a number of *points* in D and display each, together with its image in R, as we have done in Fig. 5 for the single point $z = 1 + 2i$. However, this seems hardly an attractive idea. A more appealing and commonly used device is to display the images of representative *curves*. For instance, the image of the straight line $x = 1$ $(0 < y < \infty)$ is given parametrically by $u = 1 - y^2$, $v = 2y$ or, eliminating the parameter y, by the parabola $u = 1 - v^2/4$ $(0 < v < \infty)$. Similarly, the image of $y = 1$ $(0 < x < \infty)$ is given parametrically by $u = x^2 - 1$, $v = 2x$ $(0 < x < \infty)$ or, equivalently, by the parabola $u = (v/2)^2 - 1$. And so on. The representative curves $x = 1, 2, 3, 4$, $y = 1, 2, 3, 4$, and their images are shown in Fig. 6. This type of display is of importance in Chapter 23, where we study conformal mapping. ▮

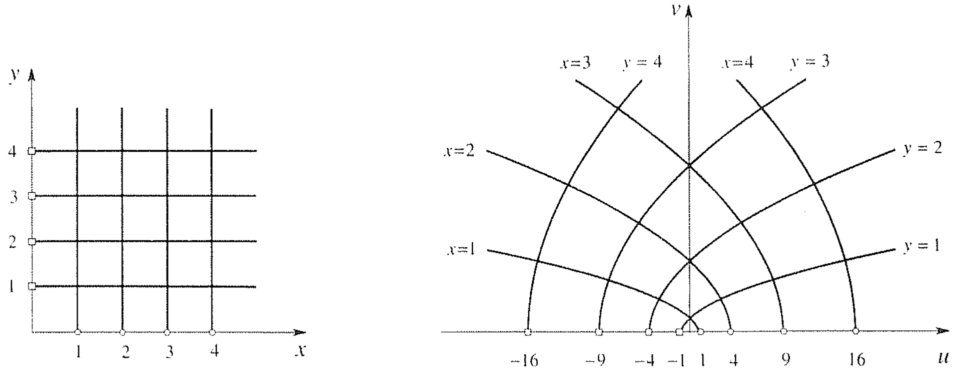


Figure 6. Representative curves and their images, under the mapping $w = z^2$.

Next, we introduce the important elementary functions $e^z, \sin z, \cos z, \sinh z$, and $\cosh z$.

21.3.2. Exponential function. Beginning with the **exponential function** e^z [also written as $\exp{(z)}$], it is essential to see, first, that we cannot "figure out" how to compute $e^z = e^{x+iy}$; the latter is a new quantity for us and we render it meaningful by *defining* it. As a rule of thumb, it is generally fruitful to define new objects in terms of, or as extensions of, old ones. For example, in the present case we could recall that for real x we have the familiar Taylor series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots . \qquad (|x| < \infty) \tag{4}$$

Thus, it seems reasonable to *define*

$$e^z \equiv 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \tag{5}$$

since (5) reduces to (4) in the event that z is purely real. In fact, this series approach was adopted by *Karl Weierstrass* (1815–1897) in his development of the complex variable theory. This approach would be awkward here since it would confront us with complex infinite series early in our development of the subject. Thus, we prefer the more traditional approach, wherein the introduction of series is delayed until after we have developed both the differential and integral calculus of functions of a complex variable.

Recalling the real variable formula $e^{x_1+x_2} = e^{x_1}e^{x_2}$, let us tentatively write

$$e^z = e^{x+iy} = e^x e^{iy}. \tag{6}$$

The latter falls short of defining e^z because e^{iy} is still undefined. At this point we do turn to Taylor series, but only to motivate our final definition of e^z. For if we treat iy as real, then (4) gives

$$\begin{aligned} e^{iy} &= 1 + (iy) + \frac{1}{2!}(iy)^2 + \frac{1}{3!}(iy)^3 + \frac{1}{4!}(iy)^4 + \cdots \\ &= \left(1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \cdots\right) + i\left(y - \frac{y^3}{3!} + \frac{y^5}{5!} - \cdots\right) \\ &= \cos y + i \sin y. \end{aligned}$$

From this result it seems reasonable to *define*

$$\boxed{e^{iy} \equiv \cos y + i \sin y} \tag{7}$$

and combining (7) with (6) to *define*

$$\boxed{e^z \equiv e^x \left(\cos y + i \sin y\right).} \tag{8}$$

Due to Euler, and known as **Euler's formula**, (7) is only a special case of (8) (for $z = iy$), but it is so important in the history of mathematics that we have framed it, together with (8). Naturally, the name of the variable in (7) is immaterial. Thus, it is equally true that $e^{i\theta} = \cos\theta + i\sin\theta$, $e^{it} = \cos t + i \sin t$, and so on.

Note carefully from (7) that

$$\left|e^{iy}\right| = |\cos y + i \sin y| = \sqrt{\cos^2 y + \sin^2 y} = 1 \tag{9}$$

for all y, and from (8) that

$$|e^z| = |e^x\left(\cos y + i \sin y\right)| = |e^x|\,|\cos y + i\sin y| = e^x \tag{10}$$

for all z. For example, $\left|e^{3-5i}\right| = e^3$ and $\left|e^{-4+i}\right| = e^{-4}$. Finally, observe from (8) that $e^z \neq 0$ for all (finite) z's because $e^x \neq 0$ for all (finite) x's, and $\cos y$ and $\sin y$ do not simultaneously vanish for any value of y. Thus, we say that e^z "has no zeros" in the finite z plane.

21.3.3. Trigonometric and hyperbolic functions. Changing y to $-y$ in (7), and recalling that the cosine is even and the sine is odd, we obtain the companion formula

$$e^{-iy} = \cos y - i \sin y, \tag{11}$$

and solving (7) and (11) for $\cos y$, $\sin y$, we obtain

$$\cos y = \frac{e^{iy} + e^{-iy}}{2}, \tag{12}$$

$$\sin y = \frac{e^{iy} - e^{-iy}}{2i}. \tag{13}$$

Equations (12) and (13) suggest that we now define the **cosine** and **sine functions** as

$$\boxed{\begin{aligned} \cos z &\equiv \frac{e^{iz} + e^{-iz}}{2}, \\ \sin z &\equiv \frac{e^{iz} - e^{-iz}}{2i}. \end{aligned}} \tag{14,15}$$

The right-hand sides are meaningful by virtue of (8). For example, $e^{iz} = e^{i(x+iy)} = e^{-y+ix} = e^{-y}(\cos x + i \sin x)$.

Furthermore, the formulas

$$\cosh x = \frac{e^{x} + e^{-x}}{2}, \tag{16}$$

$$\sinh x = \frac{e^{x} - e^{-x}}{2} \tag{17}$$

suggest that we define the **hyperbolic cosine** and **hyperbolic sine functions** as

$$\boxed{\begin{aligned} \cosh z &\equiv \frac{e^{z} + e^{-z}}{2}, \\ \sinh z &\equiv \frac{e^{z} - e^{-z}}{2}. \end{aligned}} \tag{18,19}$$

From (14), (15), (18), and (19) follow the connections

$$\cos(iz) = \cosh z, \tag{20a}$$

$$\sin(iz) = i \sinh z, \tag{20b}$$

$$\cosh(iz) = \cos z, \tag{20c}$$

$$\sinh(iz) = i \sin z \tag{20d}$$

between the trigonometric and hyperbolic functions.

Out of habit, it is natural to want to know what $\cos z$, $\sin z$, $\cosh z$, $\sinh z$ "look like." But whereas the *real* functions $\cos x, \sin x, \cosh x, \sinh x$ admit the familiar

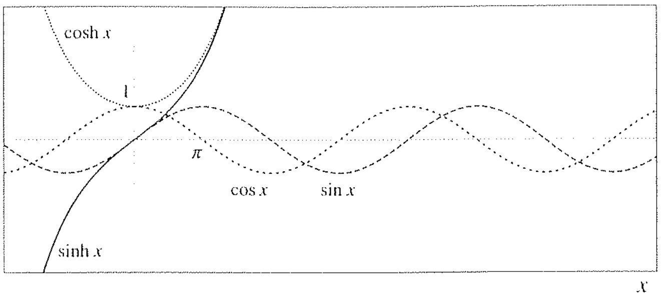


Figure 7. Graphs of $\cos x$, $\sin x$, $\cosh x$, $\sinh x$.

graphs shown in Fig. 7, recall that we cannot draw analogous graphs for the complex functions $\cos z, \sin z, \cosh z$, and $\sinh z$ because we would need four dimensions to do so. Nevertheless, a limited visual image of these functions is readily available by looking only along the real and imaginary axes. Considering $\sin z$, for example, observe that on the real axis, $\sin z = \sin x$ is the familiar oscillatory function whose graph is shown in Fig. 7. On the imaginary axis, however, $\sin z = \sin(iy) = i \sinh y$ is not oscillatory, and it is not even bounded! Incidentally, it is only by coincidence that $\sin z$ happens to be purely real along the real axis and purely imaginary along the imaginary axis. For instance, $(z+i)\sin z$ is neither purely real nor purely imaginary along either axis.

Because we have based our definitions of $e^z, \cos z, \sin z, \cosh z$, and $\sinh z$ on real variable formulas [for example, compare (18) and (19) with (16) and (17)], *it now turns out that the familiar formulas for real exponential, trigonometric, and hyperbolic functions are true for the complex functions as well.* For example,

$$\sin^2 z + \cos^2 z = 1, \qquad \sin(-z) = -\sin z, \qquad \cos(-z) = \cos z \tag{21a}$$

hold for any z, and

$$\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2, \tag{21b}$$

$$\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \sin z_2 \cos z_1 \tag{21c}$$

hold for any z_1, z_2.

EXAMPLE 2. As representative, let us prove that $\sin^2 z + \cos^2 z = 1$:

$$\begin{aligned} \sin^2 z + \cos^2 z &= -\tfrac{1}{4}\left(e^{iz} - e^{-iz}\right)^2 + \tfrac{1}{4}\left(e^{iz} + e^{-iz}\right)^2 \\ &= \tfrac{1}{4}\left(-e^{i2z} + 2 - e^{-i2z} + e^{i2z} + 2 + e^{-i2z}\right) \\ &= 1, \end{aligned} \tag{22}$$

as claimed. ▮

Additional identities are listed in the exercises.

Finally, besides $\sin z, \cos z, \sinh z, \cosh z$, we define

$$\tan z \equiv \frac{\sin z}{\cos z}, \qquad \cot z \equiv \frac{\cos z}{\sin z}, \tag{23a}$$

$$\sec z \equiv \frac{1}{\cos z}, \qquad \csc z \equiv \frac{1}{\sin z}, \tag{23b}$$

$$\tanh z \equiv \frac{\sinh z}{\cosh z}, \qquad \coth z \equiv \frac{\cosh z}{\sinh z}, \tag{23c}$$

$$\operatorname{sech} z \equiv \frac{1}{\cosh z}, \qquad \operatorname{csch} z \equiv \frac{1}{\sinh z}. \tag{23d}$$

21.3.4. Application of complex numbers to integration and the solution of differential equations. If $f(x)$ is a complex-valued function of x such as e^{ix}, then

$$\begin{aligned}\int f(x)\,dx &= \int [\mathrm{Re}f(x) + i\,\mathrm{Im}f(x)]\,dx \\ &= \int \mathrm{Re}f(x)\,dx + i\int \mathrm{Im}f(x)\,dx,\end{aligned} \tag{24}$$

where the second equality follows from the linearity of integration. Since $\int \mathrm{Re}f(x)\,dx$ and $\int \mathrm{Im}f(x)\,dx$ are real, it follows from (24) that

$$\boxed{\mathrm{Re}\int f(x)\,dx = \int \mathrm{Re}f(x)\,dx} \tag{25}$$

and

$$\boxed{\mathrm{Im}\int f(x)\,dx = \int \mathrm{Im}f(x)\,dx.} \tag{26}$$

Equations (25) and (26) can be used to simplify the evaluation of certain integrals.

EXAMPLE 3. To illustrate, let us evaluate the integral

$$I = \int_0^\infty e^{-x}\cos ax\,dx. \tag{27}$$

We can evaluate I by applying integration by parts twice, but it is simpler to proceed as follows:

$$\begin{aligned}I = \int_0^\infty e^{-x}\,\mathrm{Re}\,e^{iax}\,dx &= \int_0^\infty \mathrm{Re}(e^{-x}e^{iax})\,dx \\ &= \mathrm{Re}\int_0^\infty e^{-(1-ia)x}\,dx \quad [\text{by (25)}] \\ &= \mathrm{Re}\left(\frac{e^{-(1-ia)x}}{-(1-ia)}\right)\Bigg|_{x=0}^{x=\infty} \\ &= \mathrm{Re}\,\frac{1}{1-ia} = \frac{1}{1+a^2}.\end{aligned} \tag{28}$$

The advantage of using (25) is that the integral of $e^{-(1-ia)x}$ is simply $-e^{-(1-ia)x}/(1-ia)$, whereas the integral of $e^{-x}\cos ax$ is more difficult, requiring two applications of integration by parts.

COMMENT. Be sure to understand the reasoning behind the fifth equality:

$$\left.\frac{e^{-(1-ia)x}}{-(1-ia)}\right|_{x=0}^{x=\infty} = \lim_{x\to\infty}\frac{e^{-(1-ia)x}}{-(1-ia)} - \frac{e^0}{-(1-ia)}$$
$$= \frac{\lim_{x\to\infty}(e^{-x}e^{iax})}{-(1-ia)} + \frac{1}{1-ia}. \tag{29}$$

Further,

$$\left|e^{-x}e^{iax}\right| = \left|e^{-x}\right|\left|e^{iax}\right| = (e^{-x})(1) = e^{-x} \to 0 \tag{30}$$

as $x \to \infty$. Since the modulus of the complex number $e^{-x}e^{iax}$ tends to zero, the complex number itself must tend to zero. Thus, the first term on the right-hand side of (29) is zero and we are left with $1/(1-ia)$. ▮

Another application in which it is useful to introduce complex numbers is the determination of particular solutions of linear differential equations with cosinusoidal or sinusoidal forcing functions. For instance, consider the task of finding a particular solution $x_p(t)$ of the differential equation

$$L[x] = F\cos\omega t, \tag{31}$$

where L is a linear differential operator with constant coefficients, and F and ω are constants. According to the method of undetermined coefficients, we can obtain $x_p(t)$ in the form $C_1\cos\omega t + C_2\sin\omega t$. However, it is simpler to solve the problem

$$L[v] = Fe^{i\omega t} \tag{32}$$

for a particular solution $v_p(t)$, and to obtain x_p from v_p as

$$x_p(t) = \text{Re}\, v_p(t). \tag{33}$$

The reason the v problem is simpler is that by the method of undetermined coefficients a particular solution $v_p(t)$of (32) can be found in the form $Ae^{i\omega t}$, which form is simpler than the two-term form $C_1\cos\omega t + C_2\sin\omega t$ needed for equation (31). To verify the truth of (33), write (32) as

$$L[\text{Re}\, v + i\,\text{Im}\, v] = F\cos\omega t + iF\sin\omega t. \tag{34}$$

Since L is linear, by assumption, $L[\text{Re}\, v + i\,\text{Im}\, v] = L[\text{Re}\, v] + iL[\text{Im}\, v]$. Putting the latter expression into (34) and equating real and imaginary parts on the left- and right-hand sides gives

$$L[\text{Re}\, v] = F\cos\omega t \tag{35a}$$

and

$$L[\text{Im}\, v] = F\sin\omega t. \tag{35b}$$

Comparing (35a) with (31), the truth of (33) follows. [Of course, if the right-hand side of (31) were $F \sin \omega t$ instead, then in place of (33) we would use $x_p(t) = \operatorname{Im} v(t)$.]

EXAMPLE 4. If the applied voltage is $E(t) = E_0 \sin \omega t$, then the current $i(t)$ in the electrical circuit shown in Fig. 8 is governed by the differential equation

$$\begin{aligned} Li'' + Ri' + \frac{1}{C} i &= \frac{dE(t)}{dt} \\ &= \omega E_0 \cos \omega t, \end{aligned} \tag{36}$$

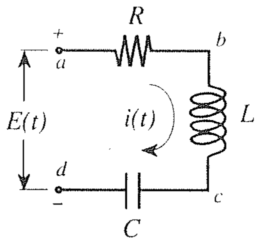


Figure 8. *RLC* circuit.

where L, R, C, E_0, ω are constants, and the inductance L is not to be confused with the operator L in (31)–(35). According to the **complex function method** described above, to find a particular solution $i_p(t)$ of (36), consider instead the simpler equation

$$Lv'' + Rv' + \frac{1}{C} v = \omega E_0 e^{i\omega t} \tag{37}$$

and seek

$$v_p(t) = Ae^{i\omega t}. \tag{38}$$

Putting (38) into (37) gives

$$\left(-L\omega^2 + iR\omega + \frac{1}{C} \right) Ae^{i\omega t} = \omega E_0 e^{i\omega t}. \tag{39}$$

Thus, $A = \omega E_0 / \left(-L\omega^2 + iR\omega + \frac{1}{C} \right)$,

$$v_p(t) = \frac{\omega E_0 C}{(1 - LC\omega^2) + iRC\omega} e^{i\omega t}, \tag{40}$$

and $i_p(t) = \operatorname{Re} v_p(t)$ gives

$$\begin{aligned} i_p(t) &= \omega E_0 C \operatorname{Re} \left(\frac{\cos \omega t + i \sin \omega t}{(1 - LC\omega^2) + iRC\omega} \, \frac{(1 - LC\omega^2) - iRC\omega}{(1 - LC\omega^2) - iRC\omega} \right) \\ &= \frac{\omega E_0 C}{(1 - LC\omega^2)^2 + R^2C^2\omega^2} [(1 - LC\omega^2) \cos \omega t - RC\omega \sin \omega t]. \end{aligned} \tag{41}$$

COMMENT 1. According to the discussion in Section 3.5, we can also express (41) in the form

$$i_p(t) = \frac{\omega E_0 C}{\sqrt{(1 - LC\omega^2)^2 + R^2C^2\omega^2}} \sin(\omega t + \phi), \tag{42a}$$

if we prefer, where the phase angle ϕ is given by

$$\phi = \tan^{-1} \left(\frac{LC\omega^2 - 1}{RC\omega} \right). \tag{42b}$$

COMMENT 2. If the inductor and capacitor were removed from the circuit, then the parentheses in (39) would contain only $iR\omega$, where R is the resistance. With the inductor and capacitor present, we can express the terms within those parentheses as

$$-L\omega^2 + iR\omega + \frac{1}{C} = i\omega\left[R + i\left(L\omega - \frac{1}{\omega C}\right)\right] \tag{43}$$

so we can think of

$$Z \equiv R + i\left(L\omega - \frac{1}{\omega C}\right) \tag{44}$$

as a sort of equivalent or generalized resistance. In electrical engineering terminology, Z is called the **complex impedance** of the circuit.

COMMENT 3. Note that because of the Ri' term in (36) the homogeneous solution of (36) will inevitably tend to zero as $t \to \infty$. Thus, (41) [or (42)] is not only a particular solution of (36), it is the **steady-state solution**. The complex function method is commonly used, in engineering, as a convenient method for obtaining the steady-state response. ▮

Closure. In this section we introduce the notion of a function w of a single complex variable z in general, and a number of elementary functions in particular. Additional elementary functions are presented in Sections 21.4 and 21.5. It would be reasonable to expect that we will subsequently turn to functions of *several* complex variables $w(z_1, \ldots, z_n)$, as one does in studying *real* variable theory. Yet it is curious and interesting that virtually the entire body of complex variable theory deals only with functions of a single complex variable. To be sure, functions of several complex variables are considered,* but this case is relatively obscure both in theory and applications.

Computer software. Graphs of the sort shown in Fig. 6 can be obtained, using *Maple*, from the **conformal** command. For instance, to obtain the image curves shown in Fig. 6, type

```
with(plots):
```

and enter, to gain access to the conformal command. Then enter

```
conformal(z^2, z = 0..4 + 4 * I, w = -20..20 + 26 * I,
grid = [5, 5], numxy = [20, 20]);
```

The $z = 0..4 + 4 * \mathrm{I}$ gives the lower left corner ($z = 0$) and upper right corner ($z = 4 + 4i$) of a rectangular region of interest in the z plane; the $w = -20..20 + 26 * \mathrm{I}$ indicates the lower left corner ($w = -20$) and upper right corner ($w = 20 + 26i$) of a rectangular region in the w plane within which the image curves are to be plotted; grid $= [m, n]$ indicates that we seek the image of m constant-x lines (the five lines $x = 0, 1, 2, 3, 4$ in this case) and n constant-y lines (the five lines $y = 0, 1, 2, 3, 4$

*See, for example, V. Vladimirov's *Methods of the Theory of Functions of Many Complex Variables* (Cambridge, MA: MIT Press, 1966), which includes applications to quantum field theory, function theory, and the theory of differential equations with constant coefficients.

in this case); and numxy $= [p, q]$ indicates the number of points to be used for each image curve. If $-20..20 + 26 * I$ were omitted, the default would be to choose the smallest rectangle, in the w plane, that contains the image. If numxy $= [20, 20]$ were omitted, the default would be numxy $= [11, 11]$, which might be too coarse to give sufficiently smooth image curves.

EXERCISES 21.3

1. With a labeled sketch, show the point sets defined by the following.

(a) $|z - 1| \leq 4$
(b) $|z + 1| < 3$
(c) $|z + 2 - i| < 2$
(d) $|z + 2 - i| = 2$
(e) $|z| < |z - 4|$
(f) $|z + 1| \leq |z|$
(g) $|z - z_1| < |z - z_2|$
(h) $\text{Re}\,(z - i) > 3$
(i) $|z + 1| = |z| + 1$
(j) $\text{Re}\,(z + i) < 2$
(k) $2 \leq |z + i| \leq 5$
(l) $\text{Im}\,(z - i) > 1$

2. Determine the range R for the given function. Include sketches of both the domain D and the range R, and give the equations of any curved parts of the boundary of R.

(a) $w(z) = z + 2 + i$ on $0 < x < 1, 0 < y < 1$
(b) $w(z) = 2iz$ on $0 < x < 1, 0 < y < \infty$
(c) $w(z) = iz + 3$ on $0 < x < \infty, 0 < y < \infty$
(d) $w(z) = z^2$ on $-\infty < x < 0, 0 < y < \infty$
(e) $w(z) = z^2$ on $1 < x < 2, 1 < y < 3$
(f) $w(z) = iz^2$ on $0 < x < 1, 0 < y < 1$
(g) $w(z) = z^3$ on $0 < x < \infty, 0 < y < \infty$

3. Show whether or not $|e^z| = e^{|z|}$. More generally, is $|w(z)| = w(|z|)$?

4. Show whether or not $\overline{e^z} = e^{\bar{z}}$. More generally, is $\overline{w(z)} = w(\bar{z})$?

5. Show that (20a) to (20d) follow from (14), (15), (18), and (19).

6. Show that

(a) $e^{z_1}e^{z_2} = e^{z_1 + z_2}$
(b) $(e^z)^n = e^{nz}$ for any integer n [More generally, it is true that $(e^{z_1})^{z_2} = e^{z_1 z_2}$, but proof of that fact needs to await our introduction of the logarithmic function in Section 21.4.]

7. Show that

(a) $\sin(-z) = -\sin z$ and $\cos(-z) = \cos z$
(b) $\cos(z_1 + z_2) = \cos z_1 \cos z_2 - \sin z_1 \sin z_2$
(c) $\sin(z_1 + z_2) = \sin z_1 \cos z_2 + \sin z_2 \cos z_1$
(d) $\cos(z + 2\pi) = \cos z$ and $\sin(z + 2\pi) = \sin z$
(e) $\cos(x + iy) = \cos x \cosh y - i \sin x \sinh y$
(f) $\sin(x + iy) = \sin x \cosh y + i \cos x \sinh y$

8. Show that

(a) $\sinh(-z) = -\sinh z$ and $\cosh(-z) = \cosh z$
(b) $\cosh(z_1 + z_2) = \cosh z_1 \cosh z_2 + \sinh z_1 \sinh z_2$
(c) $\sinh(z_1 + z_2) = \sinh z_1 \cosh z_2 + \cosh z_1 \sinh z_2$
(d) $\cosh(z + 2\pi i) = \cosh z$ and $\sinh(z + 2\pi i) = \sinh z$
(e) $\cosh(x + iy) = \cosh x \cos y + i \sinh x \sin y$
(f) $\sinh(x + iy) = \sinh x \cos y + i \cosh x \sin y$
(g) $\cosh^2 z - \sinh^2 z = 1$

9. Evaluate each of the following in standard Cartesian form.

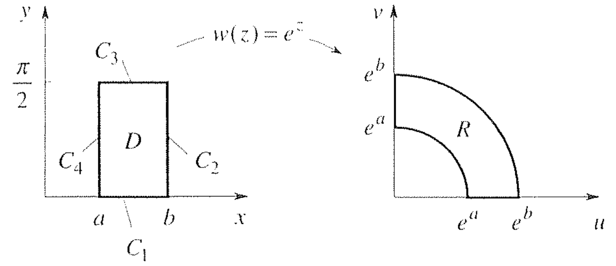
(a) $e^{2+\pi i}$
(b) e^{1-i}
(c) $e^{-\pi i/4}$
(d) $\sin(3 + \pi i)$
(e) $\cos(-2 + 3\pi i)$
(f) $\sec(1 + i)$
(g) $\csc(1 - i)$
(h) $\tan\left(-\frac{3\pi}{4}i\right)$
(i) $\cot\left(\frac{\pi i}{4}\right)$
(j) $\sinh(3 + \pi i)$
(k) $\cosh(1 - \pi i)$
(l) $\tanh(2 + 4\pi i)$

10. Consider this reasoning: $|e^{iz}| = |\cos z + i \sin z| = \sqrt{\cos^2 z + \sin^2 z} = 1$ for all z, yet $z = -2i$, say, gives $|e^{iz}| = e^2 \neq 1$. Explain the flaw(s) in that reasoning.

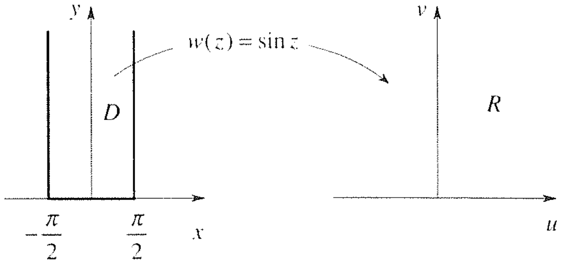
11. Prove that

(a) $e^z = 1$ if and only if $z = 2n\pi i$, where n is any integer.
(b) $e^{z_1} = e^{z_2}$ if and only if $z_1 = z_2 + 2n\pi i$, where n is any integer.

12. Show that the range R of the function e^z on the rectangular domain $a < x < b$, $0 < y < \pi/2$ is as shown in the accompanying figure. HINT: At first glance it is tempting to take a number of points on the boundary of D [such as $a, (a + b)/2, b, b + (\pi/4)i, b + (\pi/2)i$, etc.] and find their images in the f plane. That procedure can only give *points* on the boundary of R, not the boundary *curve*. Thus one should find the image of the entire line segment C_1, then C_2, C_3, and C_4.



13. Show that the range R of the function $\sin z$ on the semi-infinite strip $-\pi/2 < x < \pi/2$, $0 < y < \infty$ is shown in the accompanying figure. (See the hint in Exercise 11.)



14. Use the *Maple* conformal command, or similar software, to obtain the image of the given rectangle under the mapping $w = z^2$. Specifically, map the straight line segments comprising the border of the rectangle, as well as ten equally spaced constant-x and constant-y line segments within its interior. Label the image curves as $x =$ etc. or $y =$ etc., as we have in Fig. 6.

(a) $0 \le x \le 4,\ 0 \le y \le 4$ (b) $-4 \le x \le 4,\ 0 \le y \le 4$
(c) $2 \le x \le 4,\ 0 \le y \le 5$ (d) $2 \le x \le 4,\ 2 \le y \le 4$

15. Same as Exercise 14, but for the mapping $w = z^3$.

(a) $0 \le x \le 4,\ 0 \le y \le 4$ (b) $2 \le x \le 4,\ 0 \le y \le 4$
(c) $2 \le x \le 4,\ 2 \le y \le 6$ (d) $1 \le x \le 4,\ -1 \le y \le 0$

16. Evaluate the given integral using the complex method that is illustrated in Example 3.

(a) $\int_0^\infty e^{-x} \sin \omega x \, dx$
(b) $\int_0^\infty e^{-st} \cos \omega t \, dt \quad (\mathrm{Re}\, s > 0)$
(c) $\int_0^\infty e^{-st} \sin \omega t \, dt \quad (\mathrm{Re}\, s > 0)$

17. Show that the following statements are *incorrect*, in general.

(a) $\int_a^b \cos mx \cos nx \, dx = \mathrm{Re} \int_a^b e^{i(m+n)x} \, dx$
(b) $\int_a^b \sin mx \sin nx \, dx = \mathrm{Im} \int_a^b e^{i(m+n)x} \, dx$

18. Find a particular solution $x_p(t)$ of the given differential equation, using the complex function method that is illustrated in Example 4.

(a) $mx'' + cx' + kx = F_0 \sin \omega t$
(b) $x' + 2x = 10 \sin 3t$
(c) $x' + 3x = 20 \cos 5t$
(d) $x'' + x' = 100 \sin 5t$
(e) $x'''' + 2x' + x = 10 \sin t$
(f) $x'''' - x' + 5x = 20 \cos 2t$
(g) $x'''' - 2x' - 3x = 60 \sin 3t$
(h) $x'''' - x''' + x'' - x' + x = 25 \cos 3t$

19. The same as Exercise 12 of Section 18.4. NOTE: This exercise illustrates the use of the complex function method in solving *partial* differential equations with cosinusoidal or sinusoidal forcing functions.

21.4 Polar Form, Additional Elementary Functions, and Multi-valuedness

Having introduced the complex exponential function in the preceding section, we can now present the polar representation of complex numbers. This representation will be especially useful when, later in this section, we discuss powers of z and the logarithmic function.

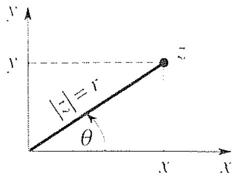


Figure 1. Polar coordinates.

21.4.1. Polar form. If r, θ are the usual polar coordinates with $r \ge 0$ and θ measured in radians counterclockwise from the positive x axis, then (Fig. 1)

$x = r\cos\theta$ and $y = r\sin\theta$, so

$$z = x + iy = r\cos\theta + ir\sin\theta = r(\cos\theta + i\sin\theta). \tag{1}$$

Recalling Euler's formula, $e^{i\theta} = \cos\theta + i\sin\theta$, we see that (1) can be re-expressed as

$$\boxed{z = re^{i\theta},} \tag{2}$$

which is the **polar** form of the complex number z. Since

$$|z| = \sqrt{x^2 + y^2} = r,$$

we see that r is the **modulus** of z:

$$r = \text{mod}\,(z) = \sqrt{x^2 + y^2}. \tag{3}$$

The angle θ, called the **argument** of z and denoted as arg (z), can be determined from the formula

$$\theta = \arg\,(z) = \tan^{-1}\frac{y}{x} \tag{4}$$

for $z \neq 0$; for $z = 0$ the angle θ is undefined, as is evident from Fig. 1. (Subsequently, we will omit the parentheses around z and will write mod z and arg z.)

Observe from Fig. 1 that given any point z ($\neq 0$) the angle θ can be determined only to within an arbitrary integer multiple of 2π. As a useful reference, the value of θ satisfying of $-\pi < \theta \leq \pi$ will be called the **principal argument** of z and will be written as Arg z. Denoting the value of Arg z as θ_0, we have

$$\theta = \arg z = \text{Arg}\, z + 2k\pi = \theta_0 + 2k\pi \tag{5}$$

for $k = 0, \pm 1, \pm 2, \ldots$. This infinite set of θ values is indeed obtained using (4) because $\tan^{-1}(\,)$ is multi-valued. However, observe carefully that (4) gives additional values as well, which are spurious and must be discarded. Let us illustrate with an example.

EXAMPLE 1. Let $z = 1 + i$, so that $x = 1$ and $y = 1$. Then (3) gives $r = \sqrt{2}$, and (4) gives $\theta = \tan^{-1}(1/1) = \tan^{-1}(1)$ which, correctly, gives $\pi/4$ plus any integer multiple of 2π. However, (4) also gives, incorrectly, $5\pi/4$ plus any integer multiple of 2π. Where do these incorrect values come from? They arise because the $\tan^{-1}(y/x)$ in (4) cannot distinguish between the given point $z = 1 + i$ ($x = y = +1$) and the point $z = -1 - i$ ($x = y = -1$), since both points give $\theta = \tan^{-1}(1)$. ▮

Observe that whereas the Cartesian form $x + iy$ is especially convenient for the addition and subtraction of complex numbers, the polar form is especially convenient for their multiplication and division. Specifically,

$$z_1 z_2 = (r_1 e^{i\theta_1})(r_2 e^{i\theta_2}) = r_1 r_2 e^{i(\theta_1 + \theta_2)}, \tag{6}$$

so that *the modulus of the product is the product of the moduli*, and *the argument of the product is the sum of the arguments*. Further, for nonzero complex numbers z_1 and z_2,

$$\frac{z_1}{z_2} = \frac{r_1 e^{i\theta_1}}{r_2 e^{i\theta_2}} = \frac{r_1}{r_2} e^{i(\theta_1 - \theta_2)} \tag{7}$$

so *the modulus of the quotient is the quotient of the moduli*, and *the argument of the quotient is the difference of the arguments*.

21.4.2. Integral powers of z and de Moivre's formula. Introducing the function z^n, where n is any integer (positive, negative, or zero), we find that its calculation is particularly convenient using the polar form of z since

$$z^n = (re^{i\theta})^n = r^n e^{in\theta}. \tag{8}$$

Equation (8) leaves the answer in polar form; if we want it re-expressed in the Cartesian form $a + ib$ we simply use the Euler formula $e^{in\theta} = \cos n\theta + i \sin n\theta$, so that

$$\boxed{z^n = r^n(\cos n\theta + i \sin n\theta),} \tag{9}$$

which result is well known as **de Moivre's formula**. There is a nagging question: Since θ is not uniquely determined, at any given point z in the complex plane, we wonder if the nonuniqueness of θ carries over to a nonuniqueness in the value of z^n. Let us see. Putting $\theta = \theta_0 + 2k\pi$ into (8) [or (9)] gives

$$z^n = r^n e^{in(\theta_0 + 2k\pi)} = r^n e^{in\theta_0} e^{i2nk\pi}. \tag{10}$$

Since $2nk$ is necessarily an even integer, $\exp{(i2nk\pi)} = \cos{(2nk\pi)} + i \sin{(2nk\pi)} = 1 + i0 = 1$ for all integers k. Thus, the nonuniqueness in θ does *not* carry over to the function z^n; z^n has the unique value $r^n \exp{(in\theta_0)}$.

EXAMPLE 2. Compute $(1+i)^3$. First, re-express $z = 1 + i$ in polar form: $r = \sqrt{2}$ and $\theta_0 = \pi/4$, so

$$(1+i)^3 = \left(\sqrt{2}e^{i\pi/4}\right)^3 = 2^{3/2} e^{i3\pi/4} \tag{11}$$

gives the result in polar form. As displayed in Fig. 2, cubing z cubes its modulus and triples its argument. If desired, we can now return to Cartesian form and express

$$(1+i)^3 = 2^{3/2}\left(\cos\frac{3\pi}{4} + i \sin\frac{3\pi}{4}\right) = 2^{3/2}\left(-\frac{1}{\sqrt{2}} + i\frac{1}{\sqrt{2}}\right) = -2 + 2i.$$

Of course, the latter result could also have been obtained by the Cartesian calculation $(1+i)^3 = (1+i)(1+i)(1+i) = (1+i)(2i) = 2i - 2$, but it is to be appreciated that the Cartesian calculation of z^n becomes quite unwieldy if n is large, whereas the polar calculation is no more difficult for large n than for small n. ▮

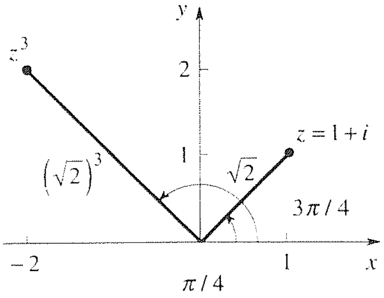


Figure 2. $1 + i$ and $(1+i)^3$.

A small technical point: in what should be the simplest case, $z = 0$, z^n is not well defined through (8) because θ is not defined at the origin. However, the

Cartesian calculation shows that z^n is zero at $z = 0$, provided that the integer n is positive.

21.4.3. Fractional powers. Next, we consider the function $z^{1/n}$, called the ***n*th root** of z. Proceeding as above, we write

$$z^{1/n} = \left(re^{i(\theta_0+2k\pi)}\right)^{1/n} = r^{1/n}e^{i(\theta_0/n)}e^{i(2k\pi/n)}, \tag{12}$$

where $k = 0, \pm 1, \pm 2, \dots$. Now, the $\exp(i2nk\pi)$ factor in (10) takes on the unique value 1 for any choice of k. Hence, as noted above, z^n takes on a unique value, namely, $r^{1/n}\exp(in\theta_0)$; that is, z^n is single-valued. However, the $\exp(i2k\pi/n)$ factor in (12) takes on n distinct values for various choices of k so that factor, and hence $z^{1/n}$ is n-valued rather than single-valued!

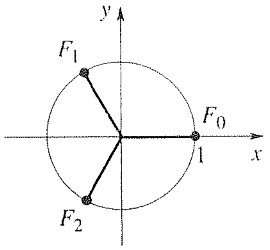


Figure 3. The three values of $\exp(i2k\pi/3)$.

EXAMPLE 3. To illustrate, consider the calculation of $(1+i)^{1/3}$. Then $r = \sqrt{2}$ and $\theta_0 = \pi/4$ so (12) gives

$$(1+i)^{1/3} = 2^{1/6}e^{i\pi/12}e^{ik(2\pi/3)}. \tag{13}$$

For $k = 0, 1, 2$, the last factor, which we will denote as F_k, takes on the values

$$F_k = e^{ik(2\pi/3)} = 1, e^{i2\pi/3}, e^{i4\pi/3}, \tag{14}$$

which correspond to the points F_0, F_1, F_2 on the unit circle shown in Fig. 3. As k increases beyond 2, the values of F_k simply repeat the values F_0, F_1, F_2 over and over; for instance, F_3 and F_4 fall upon F_0 and F_1, respectively, and F_{-1} falls upon F_2.

Putting (14) into (13) gives the three cube roots of $1+i$ as

$$(1+i)^{1/3} = 2^{1/6}e^{i\pi/12},\ 2^{1/6}e^{i9\pi/12},\ 2^{1/6}e^{i17\pi/12} \tag{15}$$

and these are depicted in Fig. 4a and denoted there as f_1, f_2, f_3. Of course, we could re-express these in Cartesian form using Euler's formula if we wish.

It is worth taking a moment to interpret this result geometrically. Consider the root f_1, for instance. Squaring f_1 squares its modulus (from $2^{1/6}$ to $2^{2/6}$) and doubles its argument (from $\theta_0/3 = \pi/12$ to $2\pi/12$), and cubing it cubes its modulus (from $2^{1/6}$ to $2^{3/6} = \sqrt{2}$) and triples its argument (from $\pi/12$ to $3\pi/12 = \pi/4$) so that f_1^3 ends up at $1+i$ as shown in Fig. 4b. Similar diagrams can be drawn for f_2 and f_3. ▮

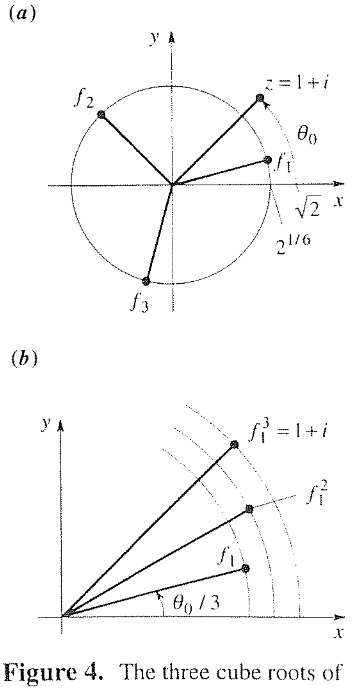


Figure 4. The three cube roots of $1+i$.

Example 3 reveals the following general pattern. The n nth roots of z (i.e., the n distinct values of $z^{1/n}$) fall on a circle of radius $|z|^{1/n}$, centered at the origin, and are equally spaced on that circle. Thus, if we can find any one root by inspection then the others fall into place on that circle. For instance, to find the three cube roots of -8 observe by inspection that one of them is -2. Then Fig. 5 shows that the other roots, equally spaced on the circle of radius 2, are $2\exp(i\pi/3)$ and $2\exp(-i\pi/3)$. You might wonder how our discussion of $z^{1/n}$ applies to $(-8)^{1/3}$ since -8 is *real*. No, we must treat it as a point in the complex plane – namely, the point $8\exp[i(\pi + 2k\pi)]$ for $k = 0, \pm 1, \pm 2, \dots$.

If we cannot find one root by inspection, then we compute r and θ_0 and use (12) with $k = 0$, say, to obtain the root $r^{1/n}e^{i(\theta_0/n)}$. With that root located, the other $n - 1$ roots are equally spaced on the circle of radius $r^{1/n}$ centered at the origin. For instance, to compute the two square roots of $1 + i$, observe that $r = \sqrt{2}$ and $\theta_0 = \pi/4$ so one root is $\sqrt{1+i} = (\sqrt{2})^{1/2}e^{i\pi/8}$, or $\left(2^{1/4}\cos\frac{\pi}{8}\right) + i\left(2^{1/4}\sin\frac{\pi}{8}\right)$. The other, being diametrically opposite this one (Fig. 6), is simply the negative of it. Similarly, $\sqrt{4} = \pm 2$ (of course) and $\sqrt{-4} = \pm 2i$.

Now that we know how to compute z^n and $z^{1/n}$, $z^{m/n}$ (where the fraction m/n is reduced to simplest form, such as $3/2$ rather than $6/4$) offers nothing new since $z^{m/n} = (z^m)^{1/n}$ is just a combination of the two and is n-valued.

EXAMPLE 4. Compute $(1 - 2i)^{7/3}$. For $z = 1 - 2i$ we have $r = \sqrt{5}$ and $\theta_0 = -1.107$ radians. Since we know that $(1 - 2i)^{7/3}$ is triple-valued, it suffices to work with the three arguments $\theta_0 = -1.107$, $\theta_0 + 2\pi = 5.176$, and $\theta_0 + 4\pi = 11.459$. Thus,

$$\begin{aligned}(1-2i)^{7/3} &= \left(\sqrt{5}e^{-1.107i}\right)^{7/3}, \quad \left(\sqrt{5}e^{5.176i}\right)^{7/3}, \quad \left(\sqrt{5}e^{11.459i}\right)^{7/3} \\ &= 6.538e^{-2.583i}, \qquad 6.538e^{12.077i}, \qquad 6.538e^{26.738i},\end{aligned}$$

which are equally spaced on the circle of radius 6.538. Naturally, these could be expressed in Cartesian form if we wish. For instance, the first is $-5.544 - 3.465i$. ∎

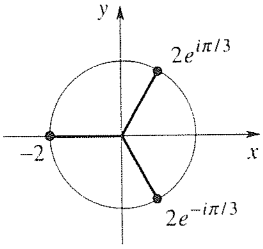


Figure 5. The cube roots of -8.

21.4.4. The logarithm of z. Next, we introduce the logarithmic function, $\log z$. To define this function, express z in polar form and write

$$\begin{aligned}\log z = \log\left(re^{i\theta}\right) &= \ln r + \log\left(e^{i\theta}\right) \\ &= \ln r + i\theta \ln e = \ln r + i\theta \\ &= \ln r + i(\theta_0 + 2\pi k),\end{aligned} \tag{16}$$

for $k = 0, \pm 1, \pm 2, \ldots$. Here, and in what follows, we distinguish the notations ln and log: we use ln to denote the ordinary logarithmic function of real variable theory (e.g., $\ln 2 = 0.693$ and $\ln e = 1$), and we use log to denote the logarithmic function of a complex variable, hereby being defined. Be sure to understand that we cannot justify the second and third equalities in (16) because they simply mimic the *real* variable theory properties $\ln(xy) = \ln x + \ln y$ and $\ln(x^y) = y \ln x$. Rather, we write (16) heuristically, and then use the result to *define*

$$\boxed{\log z \equiv \ln r + i(\theta_0 + 2\pi k)} \tag{17}$$

for $k = 0, \pm 1, \pm 2, \ldots$. Observe that just as z^n is single-valued and $z^{1/n}$ is n-valued, $\log z$ is *infinite*-valued because each choice of k produces a distinct value of $\log z$.

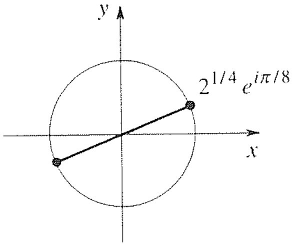


Figure 6. The square roots of $1 + i$.

EXAMPLE 5. Compute $\log(1 + i)$. With $r = \sqrt{2}$ and $\theta_0 = \pi/4$, (17) gives

$$\log(1+i) = \ln\sqrt{2} + i\left(\frac{\pi}{4} + 2\pi k\right)$$

$$= \frac{\ln 2}{2} + i\frac{\pi}{4}, \quad \frac{\ln 2}{2} + i\frac{9\pi}{4}, \quad \frac{\ln 2}{2} - i\frac{7\pi}{4}, \quad \ldots$$

for $k = 0, +1, -1$, and so on. ■

21.4.5. General powers of z. Though we have studied z^n and $z^{1/n}$ (and $z^{m/n}$ as well, where m and n are integers and m/n is reduced to simplest form), what about z^c, where c is irrational or complex? Since the polar form is so convenient for exponentiation, let us try evaluating z^c as

$$z^c = (re^{i\theta})^c = r^c e^{i(\theta_0+2\pi k)c}. \tag{18}$$

Let us not yet worry about the case where c is complex yet; consider c to be real and irrational. Then (18) does serve to give the values of z^c, and we can see that z^c is, like $\log z$, *infinite*-valued. To see that it is infinite-valued note that the factor $F_k = e^{i2\pi kc}$ starts out as $F_0 = 1$ and then takes on different values as k increases. For z^c to be finite-valued we would need F_k to return to the value 1 for *some* integer $k = K$. That will happen if and only if $2\pi Kc$ is an integer multiple of 2π, say $2\pi M$. But then $2\pi Kc = 2\pi M$ gives $c = M/K$, which is rational, whereas c was to be irrational. Thus, z^c cannot be finite-valued; it must be infinite-valued.

EXAMPLE 6. Compute $(1+i)^{\sqrt{3}}$. With $r = \sqrt{2}$ and $\theta_0 = \pi/4$, (18) gives the infinite set of distinct values

$$(1+i)^{\sqrt{3}} = (\sqrt{2})^{\sqrt{3}} e^{i(\frac{\pi}{4}+2\pi k)\sqrt{3}}$$

for $k = 0, \pm 1, \pm 2, \ldots$. For $k = 5$, for instance, we obtain $1.823e^{i41\pi\sqrt{3}/4}$ or, in Cartesian form, $1.303 - 1.275i$. ■

Finally, consider the case where c is complex: $c = a + ib$, where $b \neq 0$. In that case (18) does not help because we do not know how to evaluate the $r^c = r^{a+ib}$ factor. Thus, in place of (18) let us mimic the real variable formula $x^a = e^{\ln x^a} = e^{a \ln x}$, and *define*

$$\boxed{z^c \equiv e^{c \log z} = e^{c\,[\ln r + i(\theta_0 + 2\pi k)]}} \tag{19}$$

for $k = 0, \pm 1, \pm 2, \ldots$. Clearly, z^c is infinite-valued in this case as well since the factor $F_k = e^{i2\pi kc}$ occurs in (19), just as in (18).

EXAMPLE 7. For instance, evaluate (i.e., find all possible values of) 1^i. Here, $r = 1$ and $\theta_0 = 0$, so (19) gives

$$1^i = e^{i(0+i2\pi k)} = e^{-2\pi k},$$

for $k = 0, \pm 1, \pm 2, \ldots$. In case you thought (and not unreasonably) that the modulus of 1^i is unity, note that for $k = -20$, for instance, $1^i = e^{40\pi}$, which is enormous! ■

21.4.6. Obtaining single-valued functions by branch cuts. We have seen that functions may be single-valued (e.g., z^n), multiple-valued (e.g., $z^{1/n}$ is n-valued), or even infinite-valued (e.g., $\log z$). In some circumstances multi-valuedness is perfectly acceptable. For example, in integrating $\int dx/(x^4+1)$ by partial fractions we need to know all possible solutions of the equation $x^4+1=0$, that is, all four values of $(-1)^{1/4}$, in order to factor the x^4+1. In other cases multi-valuedness is unacceptable. For example, in the Pythagorean formula $c=\sqrt{a^2+b^2}$ we must adopt the positive square root and discard the negative square root if c is to be the *length* of the hypotenuse.

In fact, when one says that $f(z)$ is a *function* it is understood that for a given input z there is a unique output $f(z)$. Thus, when we call $z^{1/n}$ (for instance) a "multi-valued function" we abuse the understanding that functions are to be single-valued. Yet, that terminology is standard so we will not deviate from it here.

In this subsection we introduce a concept of "branch cuts" by means of which we can take multi-valued functions and render them single-valued. Consider $\log z$, for instance. We can make that function single-valued by specifying the integer k in (17). If, for instance, we choose $k=0$, then that decision amounts to restricting θ (i.e., $\arg z$) so that $-\pi<\theta\leq\pi$. Graphically, we can imagine a "slit" or "cut" in the z plane, along the negative real axis, from $x=-\infty$ all the way in to the origin, as depicted in Fig. 7a. The slit is of zero thickness; in the figure we have separated the upper and lower edges only for clarity. It is useful to regard the slit as a physical barrier that cannot be crossed. Within the *cut* plane $\log z$ is now single-valued and hence a legitimate function.

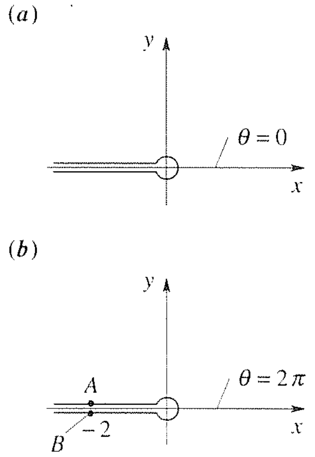


Figure 7. Two possible branch cuts for $\log z$.

The cut shown in Fig. 7a is by no means the only one possible. For instance, another suitable cut is shown in Fig. 7b, and restricts θ to the interval $\pi<\theta\leq 3\pi$. The values of the $\log z$ functions defined by Fig. 7a and b are different (at any given z point), and we can think of these two different $\log z$ functions as members, or *branches*, of a whole family of such $\log z$ functions. To illustrate, observe that if we denote them as $\log^{(a)} z$ and $\log^{(b)} z$, respectively, and use $z=i$ as representative, then $\log^{(a)} i=\pi i/2$, whereas $\log^{(b)} i=5\pi i/2$.

Suppose we are asked to compute $\log^{(b)} z$ at $z=-2$. We need to respond as follows. "Do you mean -2 on the top of the cut (point A in Fig. 7b) or on the bottom of the cut (point B)?" For the values at A and B are different: $\log^{(b)}(z_A)=\ln 2+3\pi i$, whereas $\log^{(b)}(z_B)=\ln 2+\pi i$. Understand that even though the points z_A and z_B look like "neighbors," they are actually quite far apart since to travel from z_A to z_B we need to go around the origin, since the cut is a barrier that cannot be crossed.

The arrangements shown in Fig. 7 are examples of **branch cuts**. *The role of a branch cut is to render the function single-valued.* To render $\log z=\ln r+i\theta$ single-valued we need to make θ single-valued, and we do that by introducing a cut from infinity to the origin (as a barrier against complete encirclements of the origin, which are the source of the multi-valuedness) and then defining θ at *some* point in the cut plane. For instance, in Fig. 7a we defined θ at an arbitrary point on the positive x axis. The branch of $\log z$ defined by the cut shown in Fig. 7a is known as the **principal value** of $\log z$.

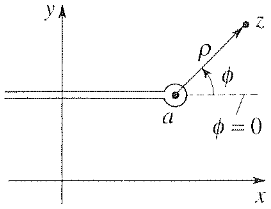


Figure 8. A branch cut for $\log(z-a)$.

How about the function $\log(z-a)$, where a is some point in the plane? The idea is to introduce a new variable, say ζ, according to $\zeta = z - a$, and to express ζ in terms of polar coordinates ρ, ϕ at that location as $\zeta = \rho e^{i\phi}$. (ζ is the "vector" from a to z.) Then

$$\log(z-a) = \log\zeta = \log(\rho e^{i\phi}) = \ln\rho + i\phi, \tag{20}$$

and if we wish to render this function single-valued then we need to introduce a branch cut such as the one shown in Fig. 8, from infinity to $\zeta = 0$ (not $z = 0$), so as to prevent encirclements of $\zeta = 0$ and the multi-valuedness that would result.

We say that $\log(z-a)$ has a **branch point** at $z = a$ because it is encirclement of that point that causes the multi-valuedness, and $\log z$ has a branch point at $z = 0$.

Similarly for other multi-valued functions. For instance, the branch cuts shown in Fig. 7 would likewise be suitable for the functions $\sqrt{z}$ and z^i, and the one shown in Fig. 8 would be suitable for $\sqrt{z-a}$ and $(z-a)^i$.

21.4.7. More about branch cuts. (Optional) To deepen our understanding of branch cuts, consider two more examples.

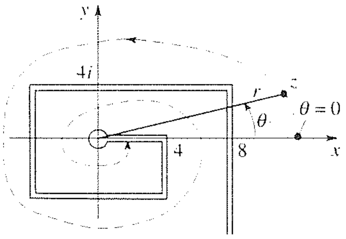


Figure 9. A more exotic branch cut.

EXAMPLE 8. It would be easy to conclude from Fig. 7 and 8, that a branch cut needs to limit θ to a 2π interval. No. To illustrate, consider the more exotic branch cut shown in Fig. 9, for the function $f(z) = z^{2/3}$. It extends straight down to $8 - i\infty$, and θ is defined as zero at the point shown. Let us evaluate $f(z)$ at $z = 2$ on the bottom of the cut. Beginning at the point where θ is zero, make a legitimate trip (i.e., one that does not cross the cut) to the point in question, as sketched in the figure. Keeping track of θ on that trip, we see that $\theta = 4\pi$ at the terminal point. Further, $r = 2$ there so

$$f(z) = z^{2/3} = (2e^{i4\pi})^{2/3} = 2^{2/3}e^{i8\pi/3}. \tag{21}$$

Making a similar trip to $z = 2$ on the *top* of the cut, from the initial point where $\theta = 0$, gives the (different) value

$$f(z) = z^{2/3} = (2e^{i2\pi})^{2/3} = 2^{2/3}e^{i4\pi/3} \tag{22}$$

there. (Do you see that $\theta = 2\pi$ there?)

The cut shown is perfectly legitimate in that it renders $z^{2/3}$ single-valued, yet it does not restrict θ to a 2π interval. Rather, $-\pi/2 < \theta \leq 4\pi$, the value $-\pi/2$ being approached as $y \to -\infty$ at any fixed x greater than 8. ▮

EXAMPLE 9. Consider one last example, the function

$$f(z) = \sqrt{z^2 - 4}. \tag{23}$$

If we write f as $\sqrt{(z-2)(z+2)}$ we see that it contains two functions, $\sqrt{z-2}$ and $\sqrt{z+2}$, each requiring its own branch cut. That is, $f(z)$ has two branch points, one at 2 and one at -2. If we wish to render f single-valued, one possible branch cut is as shown in Fig. 10a. To keep the two cuts out of each other's way we have brought one in from the right and one from the left. Each branch point serves as the origin of a polar coordinate system. To illustrate the use of this cut, let us compute f at $z = -4$ on the bottom of the cut (point c). Beginning at a, where ϕ_1 and ϕ_2 are defined to be 0, let us make the trip abc shown, keeping track of ϕ_1 and ϕ_2. Proceeding from a to b, ϕ_1 increases to π, but ϕ_2 increases and then decreases again, and equals 0 at b. Over the bc part of the trip, ϕ_1 increases a bit, but then decreases again so that $\phi_1 = \pi$ at c. Meanwhile, ϕ_2 changes from 0 at b to $-\pi$ at c. Thus, designating z at c as z_c,

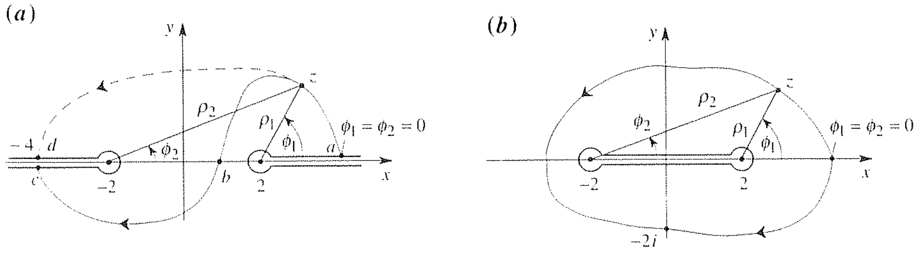


Figure 10. Branch cuts for $\sqrt{z^2 - 4}$.

$$f(z) = \left[(\rho_1 e^{i\phi_1})(\rho_2 e^{i\phi_2})\right]^{1/2} = \sqrt{\rho_1 \rho_2}\, e^{i(\phi_1+\phi_2)/2} \tag{24}$$

gives

$$f(z_c) = \sqrt{(6)(2)}\, e^{i(\pi-\pi)/2} = \sqrt{12}\, e^{i0} = \sqrt{12}. \tag{25}$$

On the *top* of the cut, at d, however, the dashed path ad gives

$$f(z_d) = \sqrt{(6)(2)}\, e^{i(\pi+\pi)/2} = \sqrt{12}\, e^{\pi i} = -\sqrt{12}. \tag{26}$$

As noted earlier, there are an infinite number of branch cuts possible for f. For instance, we could use the same cuts as in Fig. 10a, but define $\phi_1 = 4\pi$ and $\phi_2 = -26\pi$ at a. An interesting branch cut is the one shown in Fig. 10b, a *finite* cut extending from -2 to 2. This cut does *not* render ϕ_1 and ϕ_2 single-valued, yet it *does* render f single-valued, and is therefore an acceptable choice! To illustrate, let us compute $f(-2i)$. If we get to $-2i$, from the initial point (at which ϕ_1 and ϕ_2 are defined to be 0), by the counterclockwise path shown, then at $-2i$ we have $\phi_1 = 5\pi/4$ and $\phi_2 = 7\pi/4$, so (24) gives

$$f(-2i) = \sqrt{(\sqrt{8})(\sqrt{8})}\, e^{i(\frac{5\pi}{4}+\frac{7\pi}{4})/2} = -\sqrt{8}\, i. \tag{27}$$

Alternatively, if we take the clockwise path to $-2i$, shown in the figure, then $\phi_1 = -3\pi/4$ and $\phi_2 = -\pi/4$ there, so

$$f(-2i) = \sqrt{(\sqrt{8})(\sqrt{8})}\, e^{i(-\frac{3\pi}{4}-\frac{\pi}{4})/2} = -\sqrt{8}\, i$$

once again. The upshot, general proof of which is left for the exercises, is that the finite cut does indeed render f single-valued and is therefore an acceptable branch cut, even though it does not render ϕ_1 and ϕ_2 single-valued. The underlying idea is that although the finite cut permits encirclements of the branch points, it forces us to encircle *both* of them. With each encirclement of 2, $\sqrt{z-2}$ undergoes a sign change. But when we encircle 2 we also encircle -2, and the $\sqrt{z+2}$ gives another sign change, and the sign changes cancel.

Finally, there is a question as to which branch cut to choose. In the absence of any context, one branch cut is as good as another – it's your choice. However, in the context of a physical application we will see that the branch cut choice is dictated by that context. In fact, in Chapter 23 the function (23) arises in the context of a fluid mechanics application (flow over a flat plate) and we will see that the appropriate branch cut, in that case, is the one shown in Fig. 10b. See Exercise 18. ▮

Closure. The common thread in this section is the polar form of complex numbers, and its use, especially insofar as the calculation of powers of z and $\log z$. Indeed, try to evaluate $\sqrt{1+i}$, say, with*out* first reexpressing $1+i$ in polar form.

We found that z^k is single-valued only if the exponent k is an integer. If k is a rational number m/n (reduced to simplest form) it is n-valued. If k is irrational, then z^k is infinite-valued, as makes sense inasmuch as an irrational number can be expressed as the limit of a sequence of rational numbers. For instance, $\pi = 3.14159\ldots$ is the limit of the sequence $3, 31/10, 314/100 = 157/50, \ldots$ and z^3 is single-valued, $z^{31/10}$ is ten-valued, $z^{157/50}$ is 50-valued, and so on. Finally, z^k is infinite-valued also if k is complex, with $\operatorname{Im} k \neq 0$.

We note that sometimes we are interested in finding all the values of a multi-valued function, and that at other times it is necessary to render the function single-valued, as we shall see in subsequent chapters. To do so, we introduce a branch cut and define the polar angle at any specific point in the cut plane. Typically, the cut extends from the branch point to infinity, and limits the polar angle to a 2π interval. In the optional Section 21.4.7, we see that a cut of finite length can sometimes be used, and that the polar angle need not be limited to a 2π interval, nor even be made single-valued by the cut. What must always be true, however, is that the cut emanates from the branch point and renders the *function* single-valued in the cut plane.

EXERCISES 21.4

1. Determine r and the principal argument θ_0 (in radians and in degrees) for each of the following values of z.

(a) $-3i$ (b) $8i$ (c) -6
(d) $1+5i$ (e) $-4-3i$ (f) $2-12i$
(g) $-1+i$ (h) $-1-i$ (i) $0.2+i$

2. Express $\operatorname{Re}(re^{i\theta})$ and $\operatorname{Im}(re^{i\theta})$ in terms of r and θ.

3. Consider formulas (6) and (7) for the product and quotient z_1z_2 and z_1/z_2. Show that these quantities are uniquely determined even though $\arg z_1$, and $\arg z_2$ are, according to (5), multi-valued.

4. Obtain z^{10} and z^{20}, in both polar and Cartesian form, for each given z.

(a) $-1+i$ (b) $1+i$ (c) $1+2i$
(d) $2-i$ (e) $3-i$ (f) $3+4i$
(g) $5-12i$ (h) $12+5i$ (i) $2-2i$

5. Find all values of $z^{1/2}$ and $z^{1/5}$ for each given z. Express those values in polar form, and show their location in the z plane, as we have done in Fig. 5.

(a) i (b) 1 (c) 2
(d) $-i$ (e) $3-2i$ (f) $-1+2i$
(g) $3+4i$ (h) $1-4i$ (i) -32

6. Obtain, in Cartesian form, all values of $\log z$ for each given z.

(a) -2 (b) 1 (c) i
(d) $-5i$ (e) $2-i$ (f) $3-4i$
(g) $13-5i$ (h) $1+6i$ (i) $-3-2i$

7. Recall that, by definition, two complex numbers $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ are equal if $x_1 = x_2$ and $y_1 = y_2$. Show that it follows from this definition that $r_1e^{i\theta_1} = r_2e^{i\theta_2}$ (where $r_1 \neq 0$ and $r_2 \neq 0$) if $r_1 = r_2$ and $\theta_1 = \theta_2 +$ arbitrary integer multiple of 2π.

8. Obtain, in polar form, all values of $z^{2/3}, z^{3/2}$, and z^{π} for each given z.

(a) $2i$ (b) 3 (c) -4
(d) $-6i$ (e) $1-i$ (f) $2+i$
(g) $1-4i$ (h) $2+2i$ (i) $-2-2i$

9. (a)–(i) Obtain, in Cartesian form, all values of z^i and z^{1-i} for each z given in Exercise 8.

10. Let c be any real or complex number other than 0. Then for each integer value of k,

$$c^z = e^{z\log c} = e^{z[\ln|c|+i(\operatorname{Arg} c+2k\pi)]} \qquad (10.1)$$

defines a distinct single-valued function of z. Suppose that we set $k = 0$, because then if c is real and equal to e, c^z reduces to the familiar exponential function e^z. Thus let us define

$$c^z \equiv e^{z(\ln|c|+i\operatorname{Arg} c)}. \qquad (10.2)$$

We call it the **generalized exponential function** because it allows for any value of c ($\neq 0$) and reduces to the exponential function e^z for the case where $c = e$. We now state the problem. With c^z defined by (10.2), evaluate c^z for $c = 1 + \sqrt{3}\,i$ and $z = 2 - 5i$.

11. Obtain, in Cartesian form, the principal values of $\log z$ and $\sqrt{z}$ for each given z.

(a) $-3i$ (b) 2 (c) -4
(d) $2-i$ (e) $1+\sqrt{3}\,i$ (f) $-1-i$
(g) $-5i$ (h) $4-2i$ (i) $-2+4i$

12. Now that we have seen how to render $\log z$ single-valued, we can complete our discussion of the logarithmic function by stating that the familiar properties of the real logarithmic function do carry over to the complex case. For example,

$$\log(z_1z_2) = \log z_1 + \log z_2, \qquad (12.1)$$

$$\log\left(\frac{z_1}{z_2}\right) = \log z_1 - \log z_2, \qquad (12.2)$$

and

$$\log z^c = c\log z \qquad (c \text{ real or complex}), \qquad (12.3)$$

provided that all of the log functions in the equation are defined by the same branch cut. For example, (12.3) is correct if we use the principal value definition of both log functions, provided that $\arg z = \theta$ satisfies $-\pi < \theta \leq \pi$ and $\arg(z^c) = c\theta$ satisfies $-\pi < c\theta \leq \pi$.

(a) Derive (12.1). (b) Derive (12.2). (c) Derive (12.3).

13. (*Inverse of sine function*) We define the inverse of the sine function

$$w(z) = \sin^{-1} z \qquad (13.1)$$

such that $z = \sin w$.

(a) Writing the latter as

$$z = \left(e^{iw} - e^{-iw}\right)/2i,$$

show that $e^{iw} = iz + (1-z^2)^{1/2}$, and hence that

$$\sin^{-1} z = -i\log\left[iz + \sqrt{1-z^2}\right]. \qquad (13.2)$$

(b) Observe that $\sin^{-1} z$ is multi-valued because of the $(1-z^2)^{1/2}$ and also because of the log[ ]. Specifically, for each value of z ($\neq \pm 1$), the $(1-z^2)^{1/2}$ gives two values. Then, for each of these values the log gives an infinite set of values. To illustrate this point, show that

$$\sin^{-1}\left(\frac{1}{2}\right) = \frac{\pi}{6} + 2k\pi \quad \text{or} \quad \frac{5\pi}{6} + 2k\pi$$

for $k = 0, \pm 1, \pm 2, \dots$.

(c) Determine all possible values of $\sin^{-1} 2$.

(d) Determine all possible values of $\sin^{-1}(2i)$.

14. (*Inverses of other trigonometric functions*) Proceeding as in Exercise 13, derive these formulas.

(a) $\cos^{-1} z = -i\log\left[z + \sqrt{z^2-1}\right]$

(b) $\tan^{-1} z = -\dfrac{i}{2}\log\dfrac{i-z}{i+z}$

(c) $\cot^{-1} z = -\dfrac{i}{2}\log\dfrac{z+i}{z-i}$

15. (*Inverses of hyperbolic functions*) Proceeding as in Exercise 13, derive these formulas.

(a) $\sinh^{-1} z = \log\left[z + \sqrt{1+z^2}\right]$

(b) $\cosh^{-1} z = \log\left[z + \sqrt{z^2-1}\right]$

(c) $\tanh^{-1} z = \dfrac{1}{2} \log \dfrac{1+z}{1-z}$

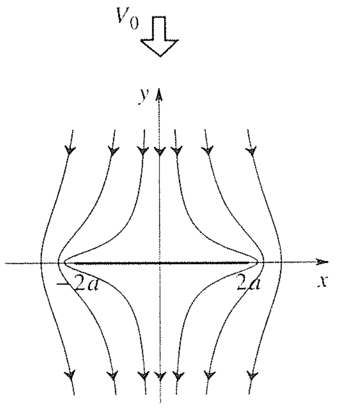
(d) $\coth^{-1} z = \dfrac{1}{2} \log \dfrac{z+1}{z-1}$

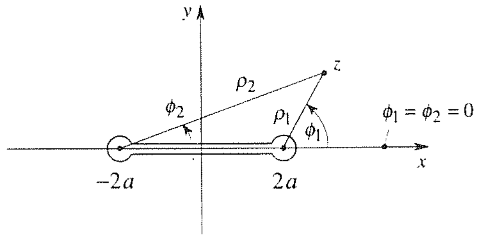
16. In Section 22.6 we will find that the plane irrotational incompressible flow of a downward free stream V_0 over a flat plate that extends from $x = -2a$ to $x = +2a$ (see the figure) is given by

$$u - iv = \frac{iV_0 z}{\sqrt{z^2 - 4a^2}}, \tag{16.1}$$

where $u(x, y)$ and $v(x, y)$ are the x and y velocity components, respectively. The context dictates using a branch cut, for $\sqrt{z^2 - 4a^2}$, extending from $-2a$ to $+2a$ (as in Example 9) as shown below. One reason for choosing that cut is that the mathematical barrier presented by the cut corresponds to the physical barrier presented by the plate. Another is that the mathematical discontinuities that arise across the cut will correspond to physically anticipated flow discontinuities across the plate. In fact, if we use a cut like the one in Fig. 10a, then those discontinuities would occur within the flow field, which would be unacceptable. The problem that we pose is for you to use (16.1) and the cut shown above to show that the velocity components on the top and bottom of the plate are

$$\begin{aligned} u(x, 0+) &= V_0 x/\sqrt{4a^2 - x^2}, \quad & v(x, 0+) &= 0, \\ u(x, 0-) &= -V_0 x/\sqrt{4a^2 - x^2}, \quad & v(x, 0-) &= 0. \end{aligned}$$





21.5 The Differential Calculus and Analyticity

Continuing to parallel the development of real variable theory, having introduced the complex number system and then a number of elementary functions defined on that number system, we next define the *limit* of a function, *continuity*, and the *derivative*. In fact, differentiability (more precisely, the closely related concept of analyticity) lies at the heart of complex variable theory and will be crucial throughout our subsequent study of conformal mapping, the complex integral calculus, and series expansions.

Let z_0 be an interior point in the domain of definition of $f(z)$. We say that *the* **limit** *of $f(z)$, as z approaches z_0, is L,*

$$\lim_{z \to z_0} f(z) = L \tag{1}$$

[or, equivalently, we write $f(z) \to L$ as $z \to z_0$] if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta > 0$ such that

$$|f(z) - L| < \epsilon \tag{2}$$

for all z's satisfying $0 < |z - z_0| < \delta$. That is (Fig. 1), we can keep the function values $f(z)$ arbitrarily close to L (namely, within the ϵ-disk) by keeping z sufficiently close to z_0 (namely, within $0 < |z - z_0| < \delta$). Observe carefully that because of the first inequality in $0 < |z - z_0| < \delta$, the value of f *at* $z = z_0$ is not at all germane. Specifically, it need not be true that $f(z_0) = L$ for (1) to hold; (1) merely states that $f(z)$ approaches L as z approaches z_0. However, if in addition to having $\lim_{z \to z_0} f(z) = L$ we also have $f(z_0) = L$, then we say that the function $f(z)$ is **continuous** at z_0.

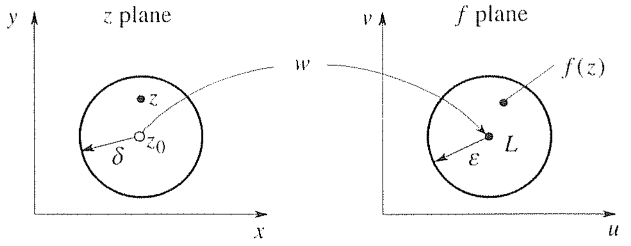


Figure 1. Limit of $f(z)$ as $z \to z_0$.

EXAMPLE 1. Show that

$$\lim_{z \to i} (z^2 + iz) = -2. \tag{3}$$

To do so, we need to show that to each $\epsilon > 0$ (no matter how small) there corresponds a $\delta > 0$ such that $|(z^2 + iz) - (-2)| < \epsilon$ for all z's in $0 < |z - i| < \delta$. No doubt, the smaller we choose ϵ, the smaller δ will need to be. Thus, $\delta = \delta(\epsilon)$, and our objective is to put forward a suitable $\delta(\epsilon)$ in order to verify (3). Since

$$\begin{aligned} |z^2 + iz + 2| &= |(z - i)(z + 2i)| = |z - i|\,|z + 2i| \\ &= |z - i|\,|(z - i) + 3i| \leq |z - i|\,(|z - i| + 3), \end{aligned} \tag{4}$$

where the last step follows from the triangle inequality, $|z^2 + iz + 2| < \epsilon$ will be satisfied if

$$|z - i|\,(|z - i| + 3) < \epsilon. \tag{5}$$

Now, solving $x(x + 3) = \epsilon$ we obtain the positive solution $x = \left(-3 + \sqrt{9 + 4\epsilon}\right)/2$ so evidently (5) will be satisfied if

$$|z - i| < \frac{-3 + \sqrt{9 + 4\epsilon}}{2}. \tag{6}$$

In summary, we have shown that given any $\epsilon > 0$ we will have $|(z^2 + iz) - (-2)| < \epsilon$ for all z's in $0 < |z - i| < \delta(\epsilon)$ if we choose

$$\delta(\epsilon) = \frac{-3 + \sqrt{9 + 4\epsilon}}{2}, \tag{7}$$

or smaller. Hence, the truth of (3) is established. In fact, it is also true that the value of $z^2 + iz$ at $z = i$ is -2 so the function $z^2 + iz$ is continuous at $z = i$. Similarly, one can show that it is continuous everywhere in the z plane. ▮

We do not wish to dwell on limits and continuity too long, since the focus of this section is differentiability. Thus, let us note merely that the functions z^n $(n = 0, 1, 2, \dots)$, $e^z, \sin z, \cos z, \sinh z$, and $\cosh z$, for example, are continuous everywhere in the z plane and the function $1/z$, for instance, is continuous for all z except $z = 0$ because $\lim_{z \to 0} 1/z$ does not exist. We are now in a position to define the derivative df/dz, which we usually denote as $f'(z)$. In doing so, we stay as close as possible to the real-variable definition. Thus, if z_0 is an interior point of the domain of definition of f, we define the **derivative** of f, at z_0, as

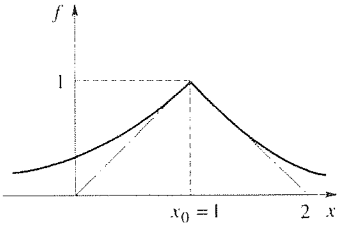


Figure 2. Approach to x_0.

$$\boxed{f'(z_0) \equiv \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z},} \tag{8}$$

provided, of course, that the limit exists. Equivalently, and sometimes more conveniently,

$$f'(z_0) \equiv \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}. \tag{9}$$

Remember that in the real-variable case, where

$$f'(x_0) \equiv \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}, \tag{10}$$

the limit value needs to be independent of the way in which x approaches x_0. For example, for the case shown in Fig. 2, $f'(x_0)$ does not exist for $x_0 = 1$ because the limiting value obtained when $x \to x_0$ from the left (namely, the slope 1) is different from the limiting value obtained when $x \to x_0$ from the right (namely, the slope -1). In fact, x does not need to approach x_0 from one side or the other. For example, the sequence $x_n = x_0 + (-1)^n e^{-n}$ $(n = 0, 1, 2, \dots)$ approaches x_0 as $n \to \infty$, but not from one side.

Analogously, in the complex case (9), we insist that the limit exist uniquely, independent of the way in which z approaches z_0. (Note that we avoid saying "independent of the direction of approach" because that wording seems to limit the mode of approach to linear as in Fig. 3a, whereas it may be spiral, for example, as in Fig. 3b. Even more so, we avoid saying "as z approaches z_0 from all directions."*

Let us illustrate the definition (8).

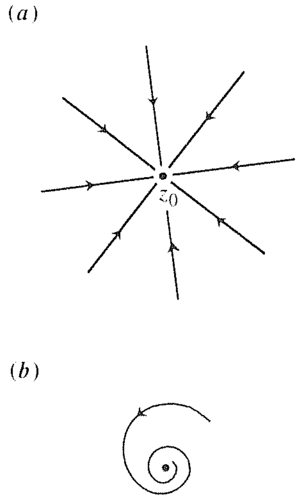


Figure 3. Approach to z_0.

EXAMPLE 2. If $f(z) = z^2$, then [dropping the subscripted zero in (8)]

$$f'(z) = \lim_{\Delta z \to 0} \frac{(z + \Delta z)^2 - z^2}{\Delta z} = \lim_{\Delta z \to 0} (2z + \Delta z) = 2z \tag{11}$$

*If this point is unclear, it should suffice to imagine Stephen Leacock's Lord Ronald, who "... said nothing; he flung himself from the room, flung himself upon his horse and rode madly off in all directions."

so that $f(z) = z^2$ is differentiable for all z, and $f'(z) = 2z$. [The last step, in (11), should be clear even if we do not go to the trouble of putting forward a $\delta(\epsilon)$ relation as we did in Example 1.] ▮

Recalling the real-variable result $d(x^2)/dx = 2x$, and the similarity between (9) and (10), it is not surprising that $d(z^2)/dz$ turned out to be $2z$. Similarly, one finds that $d(e^z)/dz = e^z$, $d(\sin z)/dz = \cos z$, and so on (Exercise 10). Furthermore, the various familiar rules of differentiation carry over to the complex case. For example, we have

$$[f(z) + g(z)]' = f'(z) + g'(z), \tag{12}$$

$$[f(z)g(z)]' = f'(z)g(z) + f(z)g'(z), \tag{13}$$

and the chain rule

$$\frac{d}{dz}f(g(z)) = f'(g(z))g'(z), \tag{14}$$

where $f'(g(z))$ denotes df/dg. Furthermore, differentiability implies continuity, and l'Hôpital's rule holds, as in the real case (Exercises 7 and 8).

From the examples cited above, one might wonder how $f(z)$ can *fail* to be differentiable.

EXAMPLE 3. $f(z) = 1/z$ is not differentiable at $z = 0$ because it is not continuous there. That is, since differentiability implies continuity, continuity is a necessary condition for differentiability. For all $z \neq 0$, however, $f'(z)$ exists (i.e., f is differentiable) and equals $-1/z^2$. ▮

EXAMPLE 4. Whereas $f(z) = 1/z$ fails to be differentiable only at a single point,

$$f(z) = \overline{z}$$

is not differentiable *anywhere*. For observe that

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{\overline{z + \Delta z} - \overline{z}}{\Delta z} = \frac{\overline{\Delta z}}{\Delta z} = \frac{\Delta x - i\Delta y}{\Delta x + i\Delta y}. \tag{15}$$

For a horizontal approach to z we have $\Delta y = 0$, and the limit of the difference quotient (15), as $\Delta x \to 0$, is $+1$. For a vertical approach, on the other hand, $\Delta x = 0$, and the limit of the difference quotient (15), as $\Delta y \to 0$, is -1. Since a unique limit (of $[f(z + \Delta z) - f(z)]/\Delta z$, as $\Delta z \to 0$) does not exist, it follows that $f(z) = \overline{z}$ is not differentiable anywhere in the z plane. ▮

Because differentiability turns out to be crucial in the study of functions of a complex variable, it will be important to find a test that can be applied to a given function $f(z) = u(x, y) + iv(x, y)$ to see if it is differentiable. Toward that end, let us first set $\Delta y = 0$ in $\Delta z = \Delta x + i\Delta y$ so that $\Delta z = \Delta x$. That is, let $z_0 + \Delta z$ approach z_0 parallel to the x axis (Fig. 4). Then the limit of the difference quotient in (8) becomes

$$\begin{aligned}\lim_{\Delta x\to 0} &\frac{[u(x_0+\Delta x, y_0) + iv(x_0+\Delta x, y_0)] - [u(x_0,y_0)+iv(x_0,y_0)]}{\Delta x} \\ &= \lim_{\Delta x\to 0}\frac{u(x_0+\Delta x,y_0)-u(x_0,y_0)}{\Delta x} + i\lim_{\Delta x\to 0}\frac{v(x_0+\Delta x,y_0)-v(x_0,y_0)}{\Delta x} \\ &= \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}. \end{aligned} \tag{16}$$

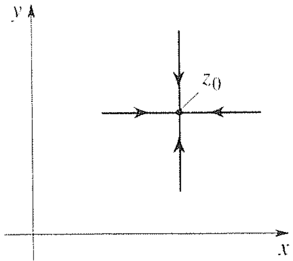


Figure 4. Horizontal and vertical approaches.

Alternatively, let us set $\Delta x = 0$ in $\Delta z = \Delta x + i\Delta y$ so that $\Delta z = i\Delta y$. That is, let $z_0 + \Delta z$ approach z_0 parallel to the y axis (Fig. 4). This time, the limit of the difference quotient in (8) becomes

$$\begin{aligned}\lim_{\Delta y\to 0} &\frac{[u(x_0, y_0+\Delta y) + iv(x_0, y_0+\Delta y)] - [u(x_0,y_0)+iv(x_0,y_0)]}{i\Delta y} \\ &= \frac{1}{i}\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}. \end{aligned} \tag{17}$$

Since the value of the limit in (8) is to be independent of the path of approach, as emphasized above, the results in (16) and (17) must agree if f is to be differentiable at z_0, that is, if $f'(z_0)$ is to exist. Thus, equating the right-hand sides of (16) and (17) shows that $u(x, y)$ and $v(x, y)$ must satisfy the relations

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}} \tag{18}$$

at the point in question. The latter are known as the **Cauchy–Riemann equations.***

Clearly, satisfaction of these equations is necessary for differentiability, but may not be sufficient since they have been obtained by considering only two possible paths of approach to the point z_0. (In fact, they are *not* sufficient, as follows from the example in Exercise 9.) We have the following theorem.

*Principal, in the development of complex variable theory were *Augustin–Louis Cauchy* (1789–1857), pronounced *ko-she*, *Georg Friedrich Bernhard Riemann* (1826–1866), and *Karl Weierstrass* (1815–1897). One of the most prolific mathematicians, Cauchy entered the Ecole Polytechnique to study engineering but, because of poor health, was advised by Lagrange and Laplace to study mathematics instead. His works span almost all branches of mathematics. Entering Göttingen to study theology, Riemann turned to mathematics, studied under Gauss, and became a professor of mathematics there in 1859. After studying law for four years at Bonn, Weierstrass likewise turned to mathematics. Weierstrass developed complex variable theory based upon the power series representation of functions, which approach failed to attract a significant following. Rather, the traditional approach, presented in this text, follows Cauchy, who began by laying a groundwork of complex differential and integral calculus. As we shall see, only after that foundation is securely laid will we be in a position to develop power series representations of functions. Thus, Weierstrass's approach was somewhat the reverse of that developed by Cauchy.

THEOREM 21.5.1 *Necessary, Sufficient Conditions for Differentiability*
Let $f(z) = u(x, y) + iv(x, y)$ be defined throughout some neighborhood of a point $z_0 = x_0 + iy_0$. For f to be differentiable at z_0,

(i) it is **necessary** that the Cauchy–Riemann equations (18) be satisfied at x_0, y_0;

(ii) it is **sufficient** that the Cauchy–Riemann equations (18) be satisfied at z_0, and that u and v be C^1 in some neighborhood of z_0.

If f is differentiable, then f' is given by any of these four equivalent expressions:

$$\boxed{f' = u_x + iv_x = v_y - iu_y = u_x - iu_y = v_y + iv_x.} \tag{19}$$

Proof: We have already proved item (i). To prove (ii), let us assume that besides the Cauchy–Riemann equations (18) being staisfied at z_0 the partial derivatives u_x, u_y, v_x, v_y are continuous in some neighborhood of z_0; in terms of the C^n notation introduced at the end of Section 13.5 we say that u and v are C^1 in some neighborhood of z_0. Consider an *arbitrary* approach to the point z_0, as indicated schematically in Fig. 5 by the curve C, and let us denote (x_0, y_0) and (x, y) as P_0 and P, respectively, for brevity. From (8),

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{[u(P) + iv(P)] - [u(P_0) + iv(P_0)]}{\Delta z}, \tag{20}$$

where Δz is the "vector" from P_0 to P. Now, since u and v have been assumed to be C^1 in some neighborhood of P_0, we can use the mean value theorem (Section 13.5) to express

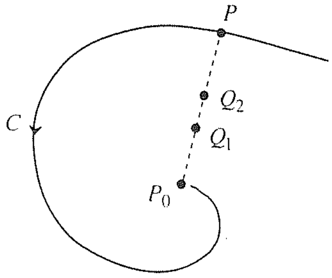


Figure 5. Arbitrary approach to z_0.

$$u(P) = u(P_0) + u_x(Q_1)\Delta x + u_y(Q_1)\Delta y \tag{21a}$$
$$v(P) = v(P_0) + v_x(Q_2)\Delta x + v_y(Q_2)\Delta y, \tag{21b}$$

where the points Q_1, Q_2 lie somewhere on the straight line between P_0 and P, as we have illustrated in Fig. 5. Using (21), we may restate (20) as

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{u_x(Q_1)\Delta x + u_y(Q_1)\Delta y + i[v_x(Q_2)\Delta x + v_y(Q_2)\Delta y]}{\Delta z}.$$

Finally, using the fact that Q_1 and Q_2 approach P_0 as $\Delta z \to 0$, let us rewrite the last result as

$$\begin{aligned} f'(z_0) &= \lim_{\Delta z \to 0} \left\{ \frac{u_x(P_0)\Delta x + u_y(P_0)\Delta y + i[v_x(P_0)\Delta x + v_y(P_0)\Delta y]}{\Delta z} + \delta \right\} \\ &= \lim_{\Delta z \to 0} \left\{ \frac{u_x(P_0)(\Delta x + i\Delta y) - iu_y(P_0)(\Delta x + i\Delta y)}{\Delta z} + \delta \right\} \\ &= u_x(P_0) - iu_y(P_0) + \lim_{\Delta z \to 0} \delta, \end{aligned} \tag{22}$$

where we have used the Cauchy–Riemann relations to obtain the second equality, and where

$$\begin{aligned}\delta = [u_x(Q_1) - u_x(P_0)]\frac{\Delta x}{\Delta z} + [u_y(Q_1) - u_y(P_0)]\frac{\Delta y}{\Delta z}\\ +i[v_x(Q_2) - v_x(P_0)]\frac{\Delta x}{\Delta z} + i[v_y(Q_2) - v_y(P_0)]\frac{\Delta y}{\Delta z}.\end{aligned} \tag{23}$$

From the continuity of the four partial derivatives, it follows that each bracketed quantity in (23) tends to zero as $\Delta z \to 0$. Furthermore, $|\Delta x| \leq |\Delta z|$ and $|\Delta y| \leq |\Delta z|$ give $|\Delta x/\Delta z| \leq 1$ and $|\Delta y/\Delta z| \leq 1$. Thus, $\lim_{\Delta z \to 0} \delta = 0$, and (22) gives the result

$$f'(z_0) = u_x(P_0) - iu_y(P_0) = u_x(x_0, y_0) - iu_y(x_0, y_0), \tag{24}$$

independent of the manner of approach of P to P_0. ■

Before continuing, observe that the additional conditions in (ii), that u_x, u_y, v_x, v_y be continuous in some neighborhood of z_0, are quite reasonable since the difference quotient limits obtained when approaching z_0 along any straight line, say, can be obtained by interpolating the limits obtained by horizontal and vertical approaches, and the validity of interpolation depends upon the continuity of the quantities being interpolated.

Let us continue. Suppose that $f(z)$ is differentiable at z_0. If, in addition, it is differentiable throughout some neighborhood of z_0, then we say that it is **analytic** at z_0. If it is not analytic at z_0, it is **singular** there, and if f is analytic at each point of a region D we say that f is analytic in D. (The terms *regular* and *holomorphic* are used by some authors, in place of analytic.)

To see how a function $f(z)$ can be differentiable at a point and yet *fail* to be differentiable in some neighborhood of the point, consider the following example.

EXAMPLE 5. Consider

$$f(z) = |z|^2 = z\bar{z}. \tag{25}$$

Then $f = (x^2 + y^2) + 0i$, so $u = x^2 + y^2$ and $v = 0$. Since

$$\begin{aligned} u_x &= 2x, \qquad & u_y &= 2y,\\ v_x &= 0, \qquad & v_y &= 0 \end{aligned}$$

we see that u, v, u_x, u_y, v_x, v_y are continuous everywhere so u and v are C^1 everywhere. Moreover, $u_x = v_y$ is satisfied at all points on $x = 0$ (the y axis) and $u_y = -v_x$ is satisfied at all points on $y = 0$ (the x axis) so both Cauchy–Riemann equations are satisfied only at the origin. Thus, $f(z) = |z|^2$ is differentiable only at $z = 0$ and hence it is analytic *nowhere*. ▮

Such behavior, however, is hardly typical of the functions that we are apt to encounter, most of which are either analytic everywhere or almost everywhere.

Functions that are analytic everywhere (i.e., at each point z in the z plane) are said to be **entire**.

EXAMPLE 6. Consider

$$f(z) = \sin z = \sin x \cosh y + i \cos x \sinh y. \tag{26}$$

Then

$$\begin{aligned} u_x &= \cos x \cosh y, & u_y &= \sin x \sinh y, \\ v_x &= -\sin x \sinh y, & v_y &= \cos x \cosh y. \end{aligned}$$

Clearly, u, v, u_x, u_y, v_x, v_y are continuous everywhere, and $u_x = v_y$ and $u_y = -v_x$ everywhere so that, according to Theorem 21.5.1, $f(z) = \sin z$ is analytic everywhere; it is an entire function. From (24),

$$f'(z) = \cos x \cosh y - i \sin x \sinh y. \tag{27}$$

To re-express the latter in terms of z we could substitute $x = (z + \overline{z})/2$ and $y = (z - \overline{z})/2i$ and simplify. The $\overline{z}$'s would cancel and we would obtain $f'(z) = \cos z$. However, in the present case it is simpler to notice, from (27), that

$$f'(z) = \cos x \cosh y - \sin x \sin (iy) = \cos (x + iy) = \cos z. \tag{28}$$

Practically speaking, we could have depended on the familiar real variable result $d(\sin x)/dx = \cos x$ to tell us that $d(\sin z)/dz = \cos z$, but the point of this example was to illustrate Theorem 21.5.1 and the concept of analyticity. ▮

Similarly, $\cos z, e^z, \sinh z, \cosh z$, and polynomial functions of z, for example, are entire functions. More complicated cases can be built from these as composite functions. Consider, for example, $f(z) = e^{\sin z}$. Recalling that

$$\frac{d}{dz} f(g(z)) = \frac{df}{dg}\frac{dg}{dz}, \quad \text{we have} \quad \frac{d}{dz} f = \left(\frac{d}{dg} e^g\right)\left(\frac{d}{dz} \sin z\right).$$

Each derivative on the right-hand side exists because the exponential and sine functions are entire; thus $f(z) = e^{\sin z}$ is entire too, and $d(e^{\sin z})/dz = e^{\sin z} \cos z$.

The function $1/z$ is analytic for all $z \neq 0$. At $z = 0$ the derivative

$$\frac{d}{dz}\left(\frac{1}{z}\right) = -\frac{1}{z^2}$$

fails to exist so $z = 0$ is a singular point of $f(z) = 1/z$. Alternatively, $z = 0$ is necessarily a singular point of $1/z$ because the latter is not continuous there.

In some cases f is expressed more conveniently in terms of the polar variables r, θ as

$$f(z) = u(r, \theta) + iv(r, \theta). \tag{29}$$

The Cauchy–Riemann equations are then found (Exercise 12) to be

$$\boxed{\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \qquad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta},} \tag{30}$$

and the derivative of f is

$$\begin{aligned} f'(z) &= e^{-i\theta}(u_r + iv_r) = \frac{e^{-i\theta}}{r}(v_\theta - iu_\theta) \\ &= e^{-i\theta}\left(u_r - \frac{i}{r}u_\theta\right) = e^{-i\theta}\left(\frac{1}{r}v_\theta + iv_r\right). \end{aligned} \tag{31}$$

EXAMPLE 7. Consider the principal value of $\log z$, defined by

$$\log z = \ln r + i\theta, \tag{32}$$

where $0 < r < \infty$ and $-\pi < \theta \leq \pi$. That is, the domain of definition of $\log z$ is the cut plane shown in Fig. 6. In this case,

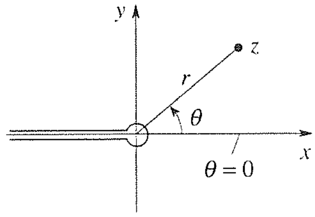


Figure 6. Domain of definition of principal value of $\log z$.

$$\begin{aligned} &u(r,\theta) = \ln r, \qquad v(r,\theta) = \theta, \\ &u_r = \frac{1}{r}, \qquad u_\theta = 0, \qquad v_r = 0, \qquad v_\theta = 1 \end{aligned} \tag{33}$$

so we see that $u, v, u_r, u_\theta, v_r, v_\theta$ are all continuous in the cut plane, and that the Cauchy–Riemann equations (30) are satisfied everywhere in the cut plane (which does not include the origin). Thus, $\log z$ is analytic everywhere in the cut plane, and (31) gives

$$\frac{d}{dz}(\log z) = e^{-i\theta}\left(\frac{1}{r} + i0\right) = \frac{1}{re^{i\theta}} = \frac{1}{z},$$

which result is no shock in view of the real-variable result $d(\ln x)/dx = 1/x$. ▮

As the final item in this section, we establish an important and interesting link between analytic function theory and partial differential equations – specifically, the Laplace equation in two dimensions. If $f(z) = u(x,y) + iv(x,y)$ is analytic in some domain D, then

$$u_x = v_y \quad \text{and} \quad u_y = -v_x \tag{34}$$

in D. Taking $\partial/\partial x$ of the first and $\partial/\partial y$ of the second,

$$u_{xx} = v_{yx} \quad \text{and} \quad u_{yy} = -v_{xy}. \tag{35}$$

We will find, in Section 23.5, that if f is analytic in D, then the partial derivatives of u and v of *all* orders exist and are continuous in D. It then follows (Section 13.3) that $v_{yx} = v_{xy}$ so that (35) gives $u_{xx} + u_{yy} = 0$ in D. Similarly, taking $\partial/\partial y$ of the first equation in (34) and $\partial/\partial x$ of the second, leads to $v_{xx} + v_{yy} = 0$. Functions that satisfy Laplace's equation and are C^2 (i.e., u, v and their various first- and second-order partial derivatives, with respect to x and y, are continuous functions of x and y) are called **harmonic functions**, so we have the following result.

THEOREM 21.5.2 *Harmonic Functions*
If $f(z) = u(x, y) + iv(x, y)$ is analytic in a domain D, then u and v are harmonic in D that is, they are C^2 and they satisfy the Laplace equations

$$\nabla^2 u = u_{xx} + u_{yy} = 0, \tag{36a}$$

$$\nabla^2 v = v_{xx} + v_{yy} = 0 \tag{36b}$$

in D.

Observe that if $f(z) = u(x, y) + iv(x, y)$ is analytic, then the harmonic functions u and v are a related pair since they are related through the Cauchy–Riemann equations. To call attention to this relationship, we refer to them as **conjugate** harmonic functions. Given one harmonic function, conjugate harmonic functions can be inferred from the Cauchy–Riemann equations.

EXAMPLE 8. It is readily verified that the function

$$u(x, y) = 3xy^2 - x^3 \tag{37}$$

is harmonic in the entire z plane. To find a conjugate harmonic function $v(x, y)$ we rely on the relations $u_x = v_y$ and $u_y = -v_x$. Thus,

$$v_x = -6xy, \tag{38a}$$

$$v_y = 3y^2 - 3x^2. \tag{38b}$$

Integrating (38a) partially, with respect to x, we obtain

$$v(x, y) = \int -6xy\, \partial x = -3x^2 y + A(y), \tag{39}$$

where $A(y)$ is arbitrary. Then, by putting (39) into (38b), we obtain

$$-3x^2 + A'(y) = 3y^2 - 3x^2,$$

whence $A(y) = y^3 + c$ for any constant c. Thus, conjugate harmonic functions corresponding to the given u are

$$v(x, y) = y^3 - 3x^2 y + c, \tag{40}$$

for arbitrary c. Further, one finds that the analytic function

$$f(z) = u + iv = (3xy^2 - x^3) + i(y^3 - 3x^2 y + c) \tag{41}$$

is actually $f(z) = -z^3 + ic$. ∎

More generally, it can be shown (Exercise 17) that given one harmonic function (u or v), a conjugate harmonic function (v or u) can always be determined (to within an arbitrary additive constant) so that $f = u + iv$ is analytic.

Closure. We begin this section by defining and discussing the differentiability of a function $f(z)$. However, differentiability at a given point will prove to be much less significant than differentiability throughout a given region, so we proceed to introduce the notion of analyticity. In a sense, the remaining chapters amount to a study of the consequences of a function's analyticity or lack of analyticity. In our closing discussion of harmonic functions we discover a connection between analytic function theory and two-dimensional potential theory – that is, between analytic function theory and the two-dimensional Laplace equation. That connection is the focus of the next chapter.

EXERCISES 21.5

1. Show that $\lim_{z\to 1} 3iz = 3i$, by referring to our definition of *limit* and actually putting forth a suitable δ. Specifically, show that $\delta = \epsilon/3$ at most. Further, show that $f(z) = 3iz$ is continuous at $z = 1$.

2. (a) Show that $\lim_{z\to 0} z^2 = 0$ by referring to our definition of *limit* and actually putting forth a suitable δ. Further, show that $f(z) = z^2$ is continuous at $z = 0$.
(b) Similarly, show that $\lim_{z\to z_0} z^2 = z_0^2$, and show that $f(z) = z^2$ is continuous for all z.

3. Prove that if $\lim_{z\to z_0} f(z) = A$ and $\lim_{z\to z_0} g(z) = B$, then
(a) $\lim_{z\to z_0}[f(z) + g(z)] = A + B$
(b) $\lim_{z\to z_0} f(z)g(z) = AB$
(c) $\lim_{z\to z_0} \dfrac{f(z)}{g(z)} = \dfrac{A}{B}$ if $B \neq 0$

4. If $f(x)$ is continuous for all (real) x, does it follow that $f(z)$ is continuous everywhere in the z plane? Explain.

5. Use the difference quotient formula (8) to evaluate $f'(z)$, as we did in Example 2.
(a) $f(z) = z^3$
(b) $f(z) = 1/z \quad (z \neq 0)$
(c) $f(z) = 1/z^2 \quad (z \neq 0)$
(d) $f(z) = 1/(z+1) \quad (z \neq -1)$

6. Use the difference quotient formula (8) to prove
(a) equation (12) (b) equation (13) (c) equation (14)

7. Prove that if $f(z)$ is differentiable at z_0, it must be continuous there. HINT: Write

$$f(z) = \frac{f(z) - f(z_0)}{z - z_0}(z - z_0) + f(z_0).$$

8. (*l'Hôpital's rule*) Prove l'Hôpital's rule, namely, that if $f(z)$ and $g(z)$ are differentiable at z_0, with $f(z_0) = g(z_0) = 0$ and $g'(z_0) \neq 0$, then

$$\lim_{z\to z_0} \frac{f(z)}{g(z)} = \frac{f'(z_0)}{g'(z_0)}. \tag{8.1}$$

9. (*Insufficiency of Cauchy–Riemann conditions*) We have seen that the Cauchy–Riemann conditions are necessary for differentiability. The example

$$f(z) = \begin{cases} \dfrac{(x^3 - y^3) + i(x^3 + y^3)}{x^2 + y^2}, & z \neq 0 \\ 0, & z = 0 \end{cases}$$

due to S. Pollard (1928), shows that they are not, however, sufficient. Verify that statement by showing that the Cauchy–Riemann conditions are satisfied at $z = 0$ but that f is not differentiable at that point.

10. Given $f(z)$, use (19) to obtain $f'(z)$. Express the answer in terms of z.

(a) $\cos z$ (b) e^{z}
(c) $\sinh z$ (d) $\cosh z$
(e) $\dfrac{1}{z}$ $(z \neq 0)$ (f) $\dfrac{1}{z+2}$ $(z \neq -2)$
(g) $\dfrac{z+1}{z-1}$ $(z \neq 1)$

11. Given $f(z)$, determine $f'(z)$, where it exists, and state where f is analytic and where it is not.

(a) $(1-2z^3)^5$ (b) $\dfrac{x+iy}{x^2+y^2}$
(c) $|z| \sin z$ (d) $\dfrac{1}{z^2+3iz-2}$
(e) $\dfrac{1}{z^3+1}$ (f) $x + i \sin y$

12. (*Cauchy–Riemann equations in polar coordinates*) Derive the Cauchy–Riemann equations (30) in the manner indicated.

(a) By carrying out the limit in (8). HINT: First let $\Delta z \to 0$ along the constant-θ line through z_0, and then let $\Delta z \to 0$ along the constant-r line through z_0. Pay careful attention to your expression for Δz in each of these cases because these cases are trickier than the cases of a horizontal approach ($\Delta z = \Delta x$) and a vertical approach ($\Delta z = i\Delta y$), used in (16) and (17).
(b) By making the change of variables $x = r\cos\theta$, $y = r\sin\theta$ in (18).

13. Determine where these functions are differentiable and where they are analytic, by checking for satisfaction of the Cauchy–Riemann equations and for continuity of u, v and their first-order partial derivatives.

(a) $f(z) = z^{100}$
(b) $f(z) = \sqrt{z}$, defined by the branch cut shown in Fig. 6
(c) $f(z) = 1/\sqrt{z}$, where the $\sqrt{z}$ is defined by the branch cut shown in Fig. 6

14. Prove that:

(a) $f(z)$ and $\overline{f(z)}$ can both be analytic, in a given region, if and only if $f(z)$ is a constant.
(b) If $f(z)$ is analytic in a region D, and $f'(z) = 0$ in D, then $f(z)$ is a constant.
(c) f cannot be analytic if it depends on $\bar{z}$. HINT: Since $f = u(x,y) + iv(x,y)$, where $x = (z+\bar{z})/2$ and $y = (z-\bar{z})/2i$, we can regard f as a function of z and $\bar{z}$. Show that the Cauchy–Riemann conditions imply that $\partial f/\partial \bar{z} = 0$, so that f must be a function of z only, if it is to be analytic. For instance, recall from Example 5 that $z\bar{z}$ is analytic nowhere.

15. Determine whether or not the given function u is harmonic and, if so, in what region. If it is, find the most general conjugate function v and corresponding analytic function $f(z)$. Express f in terms of z.

(a) $e^x \cos y$ (b) $e^{2x} \sin 2y$
(c) $x^3 - 3xy^2$ (d) $r^3 \sin 3\theta$
(e) $r^2 \cos 2\theta + 4$ (f) r
(g) $x \cos 2x \cosh 2y + y \sin 2x \sinh 2y$

16. (*Orthogonality of u = constant and v = constant curves*) (a) Prove that if $f(z) = u + iv$ is analytic in a region D, then the two families of level curves u = constant and v = constant are mutually orthogonal at all points in D at which $f'(z) \neq 0$.
(b) Illustrate the idea contained in part (a) by sketching the u and v level curves for the case $f(z) = z = x + iy$.
(c) Repeat part (b) for the case $f(z) = z^2 = (x^2 - y^2) + i2xy$.
(d) Repeat part (b) for the case

$$f(z) = \frac{1}{z} = \frac{x}{x^2+y^2} - i\frac{y}{x^2+y^2}.$$

17. (*Existence of conjugate harmonic function*) Let $u(x,y)$ be a given function that is harmonic in the rectangle $x_1 \leq x \leq x_2$, $y_1 \leq y \leq y_2$. Prove that there exists a conjugate harmonic function $v(x,y)$ in the rectangle $x_1 < x < x_2$, $y_1 < y < y_2$, say D, such that $f = u + iv$ is analytic in D. Show, further, that $v(x,y)$ is uniquely determined, to within an arbitrary additive constant. HINT: Infer from the Cauchy–Riemann equation $\partial u/\partial x = \partial v/\partial y$ that

$$v(x,y) = \int_{y_0}^{y} \frac{\partial u}{\partial x}(x,y')\partial y' + A(x),$$

and use the second Cauchy–Riemann equation and the fact that u is harmonic to obtain

$$A(x) = -\int_{x_0}^{x} \frac{\partial u}{\partial y}(x',y_0)dx',$$

where (x_0,y_0) is any point in D. Verify, with the help of the Leibniz rule (in Section 13.8), that

$$v(x,y) = \int_{y_0}^{y} \frac{\partial u}{\partial x}(x,y')\partial y' - \int_{x_0}^{x} \frac{\partial u}{\partial y}(x',y_0)dx' \tag{17.1}$$

does indeed satisfy the Cauchy – Riemann equation and is harmonic in D. Next, show that (17.1) is equivalent to the line integral representation

$$v(x,y) = \int_{(x_0,y_0)}^{(x,y)} \left[-\frac{\partial u}{\partial y}(x',y')dx' + \frac{\partial u}{\partial x}(x',y')dy'\right] \tag{17.2}$$

over any path in D, and conclude from (17.2) that v is determined only to within an arbitrary additive constant. NOTE:

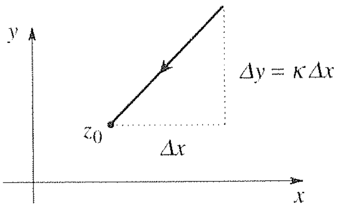
Similarly, one can show that given a harmonic function D, there exists a conjugate harmonic function u that can be determined to within an arbitrary additive constant.

18. Consider the following proposed proof of part (ii) in Theorem 21.5.1. By chain differentiation

$$\frac{df}{dz} = \frac{\partial f}{\partial x}\frac{dx}{dz} + \frac{\partial f}{\partial y}\frac{dy}{dz}. \tag{18.1}$$

If we consider a linear approach to z along a straight line with slope $dy/dx = \kappa$ (see figure), then $dz = dx + i\,dy = dx + i\kappa\,dx = (1 + i\kappa)\,dx$ so $dx/dz = 1/(1 + i\kappa)$ and $dy/dz = \kappa/(1 + i\kappa)$ in (18.1). Thus,

$$\frac{df}{dz} = (u_x + iv_x)\frac{1}{1+i\kappa} + (u_y + iv_y)\frac{\kappa}{1+i\kappa}. \tag{18.2}$$



Putting $u_y = -v_x$ and $v_y = u_x$ into (18.2), according to the Cauchy–Riemann conditions, show that (18.2) reduces to $df/dz = u_x + iv_x$ independent of κ. Then give a critical appraisal of the foregoing proposed proof, citing any weak points that you can find.

Chapter 21 Review

In this chapter we begin by introducing the complex number system, then we define the various elementary functions, and get as far as the differential calculus and analyticity.

In defining the elementary functions it is stressed that one cannot "figure out" the values of $e^z = e^{x+iy}$, for instance; e to a complex power is a new object and its values are a matter of definition. The guiding idea is to stay as close to real variable theory as possible. For instance, we write $e^z = e^{x+iy} = e^x e^{iy} = e^x[1 + iy + (iy)^2/2! + \cdots] = e^x[(1 - y^2/2! + \cdots) + i(y - y^3/3! + \cdots)] = e^x(\cos y + i \sin y)$ heuristically, and then *define* $e^z \equiv e^x(\cos y + i \sin y)$, which definition is due to Euler. Observe that on the x axis ($y = 0$), the e^z function defined in this way does agree identically with the real function e^x, and similarly for each of the other elementary functions.

Some of the elementary functions turn out to be multi-valued, namely $z^{m/n}$ (m and n integers, with m not divisible by n), $\log z$, z^c (c a real or complex number not an integer or rational number), and the inverse trigonometric and inverse hyperbolic functions. Besides showing how to determine all of their possible values, we show how to specify single-valued branches by branch cuts. A branch cut is designed to limit the domain of the function so that the function thereby defined is single-valued. Usually, the relevant polar angle(s) will be made single-valued and limited to a 2π interval, but neither of these conditions is necessary.

In the final section we define the derivative of a function $f(z)$, again staying as close to real variable theory as possible. Consequently, just as $d(\sin x)/dx = \cos x$, for instance, so does it turn out that $d(\sin z)/dz = \cos z$. The essential idea is that the limit of the difference quotient needs to exist, and have the same value, for any manner of approach to the point in question. Requiring that the limits of horizontal (parallel to the x axis) and vertical (parallel to the y axis) approach agree, gives the Cauchy–Riemann equations $u_x = v_y$ and $u_y = -v_x$ as necessary conditions for differentiability. Requiring also that the four partial derivatives be continuous in some neighborhood of the point then give sufficient conditions.

Besides differentiability at a point, we define a function to be *analytic* at z_0 if it is differentiable throughout some neighborhood of z_0. Typically, the functions that we deal with will either be analytic everywhere, or else everywhere except for one or more isolated points – singular points. Such functions will be found to be very "nice" indeed in that they are subject to powerful integral theorems, which is the subject of Chapters 23 and 24.

Chapter 22

Conformal Mapping

22.1 Introduction

In Chapter 21 we note that a function $w = f(z)$ is actually a mapping, from a given region in the z plane to a corresponding one in the w plane. However, we paid little attention to the geometrical aspects of the mapping and concentrated instead on analytical aspects such as the values of the function, differentiability, and analyticity. In the present chapter we complement that discussion by turning to the geometrical issues. Besides completing our introduction to functions of a complex variable by considering functions as mappings, the concept of conformal mapping will provide us with a powerful solution technique for problems in two-dimensional potential theory.

In Section 22.2 we lay the groundwork and in Section 22.3 we examine a particular important mapping in detail, the bilinear transformation, and explain its use in solving certain two-dimensional boundary-value problems. In Section 22.4 we give additional mappings and applications, in Section 22.5 we show how to handle more general boundary conditions, and in Section 22.6 we discuss applications specifically to fluid mechanics.

22.2 The Idea Behind Conformal Mapping

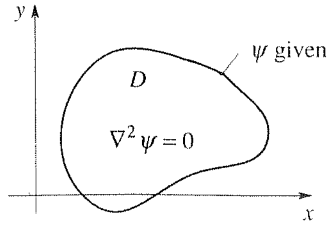


Figure 1. Generic two-dimensional potential problem with Dirichlet boundary conditions.

Suppose that we wish to solve the two-dimensional Laplace equation

$$\nabla^2 \psi = \psi_{xx} + \psi_{yy} = 0 \tag{1}$$

in some domain D of the x, y plane. We ask ψ to be C^2 in D, and to satisfy prescribed boundary conditions as well. For example, ψ might be the steady-state temperature distribution within D due to the maintaining of a certain temperature distribution along the boundary of D (Fig. 1).

The degree of difficulty in solving for ψ varies considerably with the shape of the domain D, and the methods developed in Chapter 19 fail if the shape is not sufficiently simple, such as rectangular or circular. Thus, there is interest in seeking a change of variables

$$u = u(x, y), \qquad v = v(x, y), \tag{2}$$

from x, y to u, v, in the hope that the new region (in the u, v plane), say D', will be simpler than D. Our plan, then, is to solve the new (and presumably simpler) problem in the u, v plane, and to then "return to the x, y plane" using the relations (2). Evidently, we will want a one-to-one correspondence between points in D and points in D' so that we can move back and forth unambiguously. Thus, besides asking $u(x, y)$ and $v(x, y)$ to be single valued in D, we also want single-valued *inverse* functions

$$x = x(u, v), \qquad y = y(u, v) \tag{3}$$

to exist in D'. According to the implicit function theorem (Section 13.6), we can ensure the existence of the desired inverse functions by requiring that u and v be C^1 in D and that the Jacobian be nonzero in D:

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} = u_x v_y - u_y v_x \neq 0. \tag{4}$$

With these assumptions, we proceed to enter the change of variables into (1). Denoting $\psi(x(u, v), y(u, v)) \equiv \Psi(u, v)$, chain differentiation gives

$$\psi_x = \Psi_u u_x + \Psi_v v_x$$

and

$$\psi_{xx} = (\Psi_{uu} u_x + \Psi_{uv} v_x) u_x + \Psi_u u_{xx} + (\Psi_{vu} u_x + \Psi_{vv} v_x) v_x + \Psi_v v_{xx}. \tag{5}$$

Similarly,

$$\psi_{yy} = (\Psi_{uu} u_y + \Psi_{uv} v_y) u_y + \Psi_u u_{yy} + (\Psi_{vu} u_y + \Psi_{vv} v_y) v_y + \Psi_v v_{yy} \tag{6}$$

so equation (1) becomes

$$\begin{aligned}(u_x^2 + u_y^2)\Psi_{uu} &+ (u_x v_x + u_y v_y)(\Psi_{uv} + \Psi_{vu}) \\ &+ (v_x^2 + v_y^2)\Psi_{vv} + (u_{xx} + u_{yy})\Psi_u + (v_{xx} + v_{yy})\Psi_v = 0.\end{aligned} \tag{7}$$

Thus, in making the change of variables (2) in order to simplify the domain, we inadvertently render the governing PDE much more complicated, for in place of the familiar Laplace equation (1) we have the rather unwieldy and complicated looking equation (7). For instance, since u and v are functions of x and y it follows that the coefficients in (7) are nonconstant; they are functions of x and y which, through the inverse relations (3), are functions of u and v.

The result seems fair enough: if we make a change of variables (2) to simplify the region D, we can expect to complicate the PDE. However, it is striking that if we restrict the change of variables (2) so that $u(x,y)$ and $v(x,y)$ satisfy the relations

$$u_x = v_y \quad \text{and} \quad u_y = -v_x \tag{8}$$

in D, and ask u and v to be C^2 so that $u_{xy} = u_{yx}$ and $v_{xy} = v_{yx}$ in D, then (8) gives $u_x v_x + u_y v_y = (u_x)(-u_y) + (u_y)(u_x) = 0$, $u_{xx} + u_{yy} = v_{yx} - v_{xy} = 0$, and $v_{xx} + v_{yy} = -u_{yx} + u_{xy} = 0$ so (7) simplifies dramatically to

$$(u_x^2 + u_y^2)(\Psi_{uu} + \Psi_{vv}) = 0. \tag{9}$$

Using (8) again, we can write $u_x^2 + u_y^2 = u_x v_y - u_y v_x$, which Jacobian has already been assumed nonzero in D. Then it follows from (9) that

$$\boxed{\Psi_{uu} + \Psi_{vv} = 0} \tag{10}$$

everywhere in D'; that is, the Laplace equation is preserved under the change of variables!

Notice carefully that thus far there has been no mention, or use, of complex variable theory. However, the fact that equations (8) are the familiar Cauchy–Riemann conditions causes us to raise an eyebrow, and suggests that it might be helpful (though not essential) to regard the x, y and u, v planes as complex z and w planes, where $z = x + iy$ and $w = f(z) = u(x,y) + iv(x,y)$. (Whether the x, y and u, v planes are real planes or complex planes is merely a matter of viewpoint.) Then $u_x v_y - u_y v_x = u_x^2 + v_x^2 = |f'(z)|^2$, so the nonvanishing-Jacobian condition (4) can be stated more crisply in terms of complex variable theory as $f'(z) \neq 0$ in D, and we have established the following important result.

THEOREM 22.2.1 *Preservation of Laplace Equation*
Let $w = f(z) = u(x,y) + iv(x,y)$ be analytic everywhere in a domain D, with $f'(z) \neq 0$ everywhere in D. Denote the image of D as D', and let the mapping be one-to-one. If $\psi(x,y)$ is harmonic in D, then $\psi(x(u,v), y(u,v)) \equiv \Psi(u,v)$ is harmonic in D'.

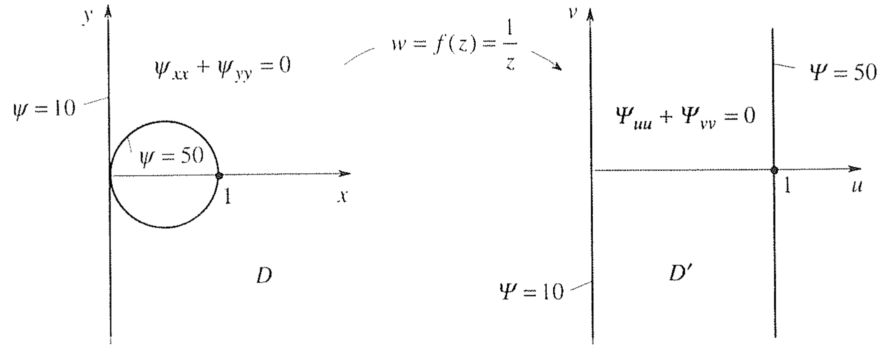
The point, then, is that if we use an analytic function $f(z) = u(x,y) + iv(x,y)$ to give the change of variables (2), then the Laplace equation on $\psi(x,y)$ is guaranteed to carry over to a Laplace equation on $\Psi(u,v)$. At the same time, there is the possibility of selecting the analytic function $f(z)$ so as to simplify the region.

Let us make two remarks about the theorem. First, recall that D is open because it is a domain. Thus, $f'(z) \neq 0$ in D means at all interior points of D; $f'(z)$ need not be nonzero on the boundary of D. Put differently, we ask that the Laplace equation $\nabla^2\psi = 0$ be preserved only on the interior of D because (as noted in the chapters on PDE's) we require the satisfaction of that equation only on the interior of D. Second, why do we change notation, from ψ to Ψ? Because ψ and Ψ are different functions. For instance, suppose that $\Psi(u,v) = u + 2v$

and $f(z) = z^2$. From $f(z) = z^2$ it follows that $u = x^2 - y^2$ and $v = 2xy$ so $\Psi(u,v) = u + 2v = x^2 - y^2 + 4xy = \psi(x,y)$, and surely $\Psi(u,v) = u + 2v$ and $\psi(x,y) = x^2 - y^2 + 4xy$ are different functions; for instance, $\Psi(2,3) = 8$ whereas $\psi(2,3) = 19$.

We call the function $f(z)$ a **conformal map**. The relevance of the adjective conformal will be clarified by the next theorem. However, before proceeding with that let us immediately give an application, to show how the method is used.

EXAMPLE 1. *Application of Theorem 22.2.1.* Consider the Dirchlet problem shown in Fig. 2: the Laplace equation on $\psi(x,y)$ together with Dirichlet boundary conditions that $\psi = 50$ on the circular part of the boundary and $\psi = 10$ on the straight part, and that ψ is bounded on D. Physically, $\psi(x,y)$ might be the steady-state temperature distribution in a large plate with a circular cutout near one edge.



y
$\psi_{xx} + \psi_{yy} = 0$
$\psi = 10$
$\psi = 50$
1
x
D
$w = f(z) = \dfrac{1}{z}$
v
$\Psi = 50$
$\Psi_{uu} + \Psi_{vv} = 0$
1
u
D'
$\Psi = 10$

Figure 2. Steady-state temperature problem.

First, observe that the given problem cannot be solved by separation of variables in Cartesian coordinates (because the boundary is not comprised of constant-x and constant-y curves), nor in plane polar coordinates (because no matter where the origin is located the boundary is not comprised of constant-r and constant-θ curves). Nor do the Laplace and Fourier transforms provide any help.

Turning to Theorem 22.2.1, suppose we know that the mapping $f(z) = 1/z$ sends the given domain D into the simpler strip D' shown in Fig. 2 (Exercise 1). Furthermore, $f(z) = 1/z$ is analytic for all $z \neq 0$, and hence everywhere inside D; $f'(z) = -1/z^2 \neq 0$ everywhere in D, and the mapping is one-to-one since both $f(z) = 1/z$ and the inverse function $z = 1/f$ are, obviously, single-valued. Thus, it follows from Theorem 22.2.1 that

$$\nabla^2\Psi = \Psi_{uu} + \Psi_{vv} = 0 \tag{11}$$

in D'. Finally, the images of the boundary curves $x = 0$ and $(x - \frac{1}{2})^2 + y^2 = \frac{1}{4}$ are $u = 0$ and $u = 1$ so that $\Psi(0,v) = 10$ and $\Psi(1,v) = 50$ as indicated in Fig. 2. And since ψ is to be bounded on D, Ψ must be bounded on D'.

Observe that whereas the ψ problem is hard, the Ψ problem is easy because (see Fig. 2) it is evidently one-dimensional. That is, Ψ varies with u but not v. Then the Laplace equation $\Psi_{uu} + \Psi_{vv} = 0$ reduces to the ordinary differential equation $d^2\Psi/du^2 = 0$, with general solution $\Psi = Au + B$. The boundary conditions $\Psi(0, v) = 10$ and $\Psi(1, v) = 50$ give $A = 40$ and $B = 10$ so

$$\Psi(u, v) = 10 + 40u \tag{12}$$

in D'. To obtain the solution for $\psi(x, y)$ in D, we merely substitute $u = u(x, y)$ and $v = v(x, y)$ into the right-hand side of (12). (In this case there happen to be no v's.) We get $u(x, y)$ and $v(x, y)$ from the mapping f. Specifically, since $f(z) = 1/z = u + iv$, we have

$$u + iv = \frac{1}{x + iy}\frac{x - iy}{x - iy} = \frac{x}{x^2 + y^2} - i\frac{y}{x^2 + y^2}$$

so

$$u = \frac{x}{x^2 + y^2}, \qquad v = -\frac{y}{x^2 + y^2}. \tag{13}$$

Thus,

$$\psi(x, y) = \Psi(u(x, y), v(x, y)) = 10 + \frac{40x}{x^2 + y^2}. \tag{14}$$

COMMENT 1. Notice that whereas the method of separation of variables, used extensively in Chapters 18–20, generally gives the solution as an infinite series, (14) is expressed, more conveniently, in closed form.

COMMENT 2. In case it was not clear that $\Psi(u, v)$ is independent of v, as claimed, observe a posteriori that (12) does indeed satisfy the full equation (11) as well as the boundary conditions. The only nagging question is whether or not the solution (12) is *unique*, for perhaps there are other solutions, as well, that do vary with v. This is a subtle point, and is discussed in Exercise 1.

COMMENT 3. We stated, without explanation, that the boundary conditions carry over, from the z plane to the w plane. For example, if $\psi = 20$ on some boundary curve in the z plane, then $\Psi = 20$ on the image of that curve in the w plane. That result follows immediately from the fact that $\psi(x, y) = \psi(x(u, v), y(u, v)) = \Psi(u, v)$. That is, if the point (u, v) (or $u + iv$ in complex notation) is the image of (x, y) (i.e., $x + iy$), then Ψ at (u, v) is equal to ψ at (x, y). ▮

Though the title of this chapter is Conformal Mapping, we have not yet introduced the notion of conformality. Recall that Theorem 22.2.1 requires that $f(z)$ be analytic and that $f'(z) \neq 0$ in the given domain. As we shall prove, such mappings are "conformal." That is, they preserve angles both in magnitude and in sense. To make this claim precise, consider two oriented smooth curves C_1 and C_2 that intersect at z_0 (Fig. 3). Let C_1 and C_2 be parametrized by $z_1(\tau) = x_1(\tau) + iy_1(\tau)$ and $z_2(\tau) = x_2(\tau) + iy_2(\tau)$, respectively, where x_1, y_1, x_2, y_2 are differentiable functions of the real parameter τ, where $z_1(0) = z_2(0) = z_0$ and where $\dot{z}_1(0)$ and $\dot{z}_2(0)$ are both nonzero. (Dots will be used to denote $d/d\tau$.) Then the complex number

$$\dot{z}_1(0) = \lim_{\tau \to 0} \frac{z_1(\tau) - z_1(0)}{\tau} = \dot{x}_1(0) + i\dot{y}_1(0) \tag{15}$$

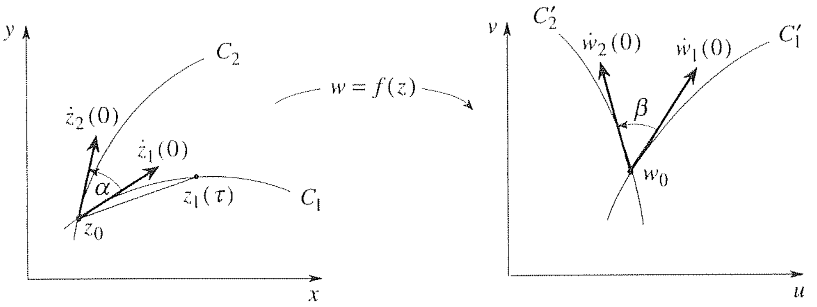


Figure 3. Conformality.

is a tangent "vector" to C_1 at z_0 and similarly for $\dot{z}_2(0)$. Let the angle measured counterclockwise from $\dot{z}_1(0)$ to $\dot{z}_2(0)$ be denoted as α.

Now, consider a mapping $w = f(z)$ which is a nonconstant analytic function. Under this mapping, z_0 is sent into $w_0 = f(z_0)$ and the curves C_1, C_2 are sent into curves C_1', C_2' that are parametrized by $w_1(\tau) = f(z_1(\tau))$, $w_2(\tau) = f(z_2(\tau))$, respectively. By chain differentiation,

$$\dot{w}_1(0) = f'(z_0)\dot{z}_1(0), \qquad \dot{w}_2(0) = f'(z_0)\dot{z}_2(0) \tag{16}$$

so that if (and only if) $f'(z_0) \neq 0$, there exist unique tangent vectors $\dot{w}_1(0), \dot{w}_2(0)$, at w_0, given by (16) and depicted in Fig. 3. From (16) it follows that

$$\begin{aligned} \arg \dot{w}_1(0) &= \arg f'(z_0) + \arg \dot{z}_1(0), \\ \arg \dot{w}_2(0) &= \arg f'(z_0) + \arg \dot{z}_2(0), \end{aligned} \tag{17}$$

and subtraction gives

$$\arg \dot{w}_1(0) - \arg \dot{w}_2(0) = \arg \dot{z}_1(0) - \arg \dot{z}_2(0). \tag{18}$$

If we denote the angle $\arg \dot{w}_1(0) - \arg \dot{w}_2(0)$ as β (Fig. 3), (18) tells us that

$$\beta = \alpha, \tag{19}$$

and this result is the promised conformality. (See Exercise 8.) Observe that the result (19) holds for *any* two oriented smooth curves C_1, C_2 intersecting at z_0. Let us pull these results together now by defining conformality and then giving conditions that ensure that a mapping is conformal.

We say that a mapping $w = f(z)$ is **conformal** at z_0 if it preserves, both in magnitude and in sense, the angle between every pair of oriented smooth curves that intersect at z_0.

THEOREM 22.2.2 *Conformality*
If $f(z)$ is analytic, then the mapping $w = f(z)$ is conformal, except at points where $f'(z) = 0$.

EXAMPLE 2. *Illustration of Conformality.* To emphasize the *local* nature of conformality, consider the mapping $w = z^2$. If the domain of definition D is the first quadrant $0 < x < \infty$, $0 < y < \infty$, then the range D' is the upper half plane $v > 0$ as shown in Fig. 4. In this case f is analytic everywhere, and $f'(z) = 2z = 0$ only at the origin. Thus, the mapping $w = z^2$ is conformal everywhere in D. If angles are preserved, we may well wonder how the quarter plane manages to open out into the upper half plane. Thus, consider, as representative, the curve ABC and its image $A'B'C'$. The angle of intersection at B, namely $\pi/2$, is indeed preserved, although *away from the neighborhood of* B the curve ABC opens out like the pages of an open book (Fig. 4). Thus, conformality is local, not global. At $z = 0$, the only point at which conformality breaks down, the angle $\pi/2$ is *not* preserved; it is doubled.

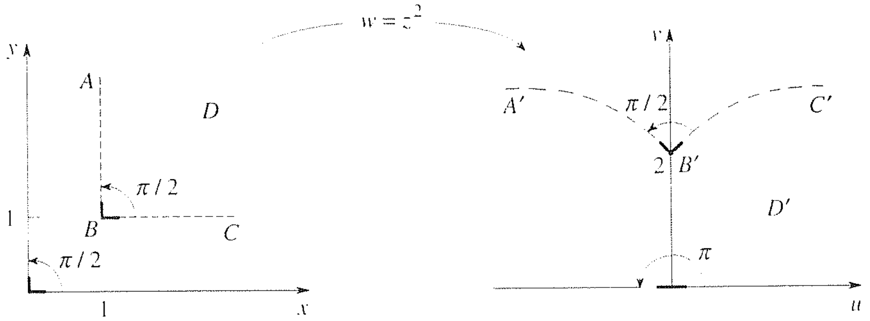


Figure 4. The mapping $w = z^2$.

COMMENT. We do not mean to imply that if conformality breaks down at a point, then angles are necessarily doubled there. That result is specific to the present example. ▮

Closure. We show first that under mappings $w = f(z)$, where f is analytic and $f'(z) \neq 0$, the Laplace equation is preserved. This fact is of great importance in two-dimensional potential theory for it implies that given a particular problem in two-dimensional potential theory, one might be able to effect a simplification in the shape of the domain without disturbing the governing Laplace equation. We illustrate this line of approach in Example 1, and plan to develop it more fully in the sections to follow. Furthermore, we show that the above-noted conditions on f guarantee conformality of the map.

Some additional results are included in the exercises, but important questions remain. For a given domain D and a desired image D', does a one-to-one conformal map *exist*? If so, is it *unique*? These questions are answered, for simply connected domains (other than the entire z plane), by the **Riemann mapping theorem.**[*] The existence part of the theorem states, in effect, that any simply connected domain D (not the entire plane) can be mapped one-to-one and conformally onto any other simply connected domain D' (not the entire plane).

EXERCISES 22.2

1. (*More about Example 1*) (a) Show that the image of the domain D shown in Fig. 2, under the mapping $w = f(z) = 1/z$, is the infinite strip D' shown in the figure.
(b) Using separation of variables, show that the problem $\Psi_{uu} + \Psi_{vv} = 0$ in the strip $0 < u < 1$ and $-\infty < v < \infty$, with $\Psi(0, v) = 10$ and $\Psi(1, v) = 50$, admits the solution

$$\Psi(u, v) = 10 + 40u + (Ae^{\pi v} + Be^{-\pi v}) \sin \pi u, \qquad (1.1)$$

where A, B are arbitrary constants. NOTE: Observe that if we fail to also require that ψ, and hence Ψ, be bounded, then the solution is nonunique because A and B are arbitrary in (1.1). If we do require that Ψ be bounded, then we need $A = 0$ because $e^{\pi v} \to \infty$ as $v \to \infty$, and we need $B = 0$ because $e^{-\pi v} \to \infty$ as $v \to -\infty$, so (1.1) reduces to $\Psi(u, v) = 10 + 40u$ as given by (12). Although we see from (1.1) that the solution is not unique if a boundedness condition is not included, understand that we have not proved uniqueness if a boundedness condition *is* included because there might be solutions of $\nabla^2\Psi = 0$, besides (1.1), that cannot be obtained by the method of separation of variables.
(c) If A and B are not both zero in (1.1), then Ψ grows unboundedly as $|v| \to \infty$. What is the corresponding behavior of ψ in the x, y plane?

2. (*Analyticity of inverse mapping*) If $w = f(z)$ is a one-to-one mapping, there exists a single-valued inverse function, say $z = f^{-1}(w)$. Show that if f is an analytic function of z, then f^{-1} (i.e., the inverse function, not the numerical inverse $1/f$) is an analytic function of w, and

$$\frac{df^{-1}(w)}{dw} = \frac{1}{df(z)/dz}$$

or, equivalently,

$$\frac{dz}{dw} = \frac{1}{dw/dz}. \qquad (2.1)$$

HINT: Show that if $\Delta z \to 0$, then $\Delta w \to 0$, too.

3. Give a shorter proof of Theorem 22.2.1 using the fact that the inverse mapping $z(w)$ is analytic, too, as noted in the preceding exercise. HINT: There exists a conjugate harmonic function corresponding to $\psi(x, y)$, say $\zeta(x, y)$, such that

$$F(z) = \psi(x, y) + i\zeta(x, y) \qquad (3.1)$$

is an analytic function of z. Thus

$$\begin{aligned} F(z(w)) &= \psi(x(u, v), y(u, v)) + i\zeta(x(u, v), y(u, v)) \\ &= \Psi(u, v) + iZ(u, v). \end{aligned} \qquad (3.2)$$

Show why it follows from (3.2) that $\Psi(u, v)$ is harmonic.

4. (*Local magnification*) Let $w = f(z)$ be analytic at z_0, with $f'(z_0) \neq 0$. Denoting $\lim_{\Delta z \to 0} |\Delta w/\Delta z|$ as the *local magnification*, show that the local magnification is equal to $|f'(z_0)|$.

5. Are the following mappings conformal at $z = 0$? If not, why not?

(a) e^z (b) ze^z (c) $iz + 3$
(d) iz^2 (e) $\sin z$ (f) $1/(z - 1)$
(g) $z^2 + z$ (h) $z^3 - 1 + 2i$

6. (a) In Example 2 we observed that the mapping $w = z^2$ is conformal for all z except $z = 0$. At $z = 0$ we noted that the "corner angle" $\pi/2$ gets doubled to π. Show that *all* angles are doubled at $z = 0$ (i.e., between every pair of oriented smooth curves that intersect at $z = 0$).
(b) More generally, show that for the mapping $w = z^n$ for any integer $n = 2, 3, 4, \ldots$, angles are multiplied by n at $z = 0$.
(c) Still more generally, suppose that $w = f(z) = (z - z_0)^n g(z)$, where $g(z)$ is analytic, $g(z_0) \neq 0$, $g'(z_0) \neq 0$, and n is any positive integer greater than 1. Show that angles are multiplied by n at z_0; i.e., in place of (19) we have $\beta = n\alpha$.

[*] Riemann's celebrated mapping theorem appears near the end of his doctoral dissertation. For discussion of this theorem, see E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis* (Englewood Cliffs, NJ: Prentice Hall, 1976).

7. Let $w = f(z) = u(x,y) + iv(x,y)$ be analytic everywhere in a domain D, with $f'(z) \neq 0$ everywhere in D. Denote the image of D as D', and let the mapping be one-to-one. If $\psi(x,y)$ satisfies the *Poisson equation*

$$\psi_{xx} + \psi_{yy} = Q(x,y),$$

where Q is prescribed, what equation does $\psi(x(u,v), y(u,v)) = \Psi(u,v)$ satisfy in D'?

8. (*A simpler proof of conformality*) In case our proof of conformality [equation (19)] is unclear, let us suggest a simplified version of that proof, even though it is only heuristic by virtue of its use of differentials as computable quantities. Let dz_1 and dz_2 be infinitesimal "vectors" springing from z_0, and let dw_1 and dw_2 be their respective image vectors springing from w_0. Writing $dw_2 = f'(z_0)dz_2$ and $dw_1 = f'(z_0)dz_1$, conclude that $\arg dw_2 = \arg f'(z_0) + \arg dz_2$, and that $\arg dw_1 = \arg f'(z_0) + \arg dz_1$. Show that (19) follows from these two equations. Include a labeled sketch in your solution.

9. Here is a subtle question regarding Theorem 22.2.1. If the condition $f'(z) \neq 0$ ensures the existence of single-valued inverse functions $x = x(u,v)$ and $y = y(u,v)$, and the "forward" mapping $f(z)$ is of course single valued, why do we bother including the additional condition that "the mapping be one-to-one" in the theorem?

10. What is equation (7) for the case where $x = u \cos v$ and $y = u \sin v$ (i.e., u and v are really the polar coordinates r and θ)? Is the map

$$w(z) = u + iv = \sqrt{x^2 + y^2} + i \tan^{-1} \frac{y}{x}$$

conformal? Explain.

22.3 The Bilinear Transformation

The transformation $w(z) = 1/z$ that was used in Example 1 of the preceding section is a special case of the **bilinear transformation**[*]

$$\boxed{w = \frac{az + b}{cz + d} \qquad \text{for } ad - bc \neq 0,} \tag{1}$$

which is also known as the *Möbius transformation* or *linear fractional transformation*. The constants a, b, c, d are allowed to be complex, and the condition $ad - bc \neq 0$ ensures that $w(z)$ is not merely a constant, or $0/0$.

Since

$$w'(z) = \frac{ad - bc}{(cz + d)^2} \tag{2}$$

exists for all z except $z = -d/c$ (at which point w is undefined as well) and is nonzero (since $ad - bc \neq 0$ by assumption), we see that *the bilinear transformation is conformal everywhere except at $z = -d/c$*, which point is called the **pole** of the transformation. Although w is undefined at $z = -d/c$, it will be convenient to regard the image of that point as the **point at infinity** in the w plane, denoted as $w = \infty$. To indicate that the w plane has been augmented by the point at infinity, we call the result the **extended complex plane**. In contrast, the "usual" complex plane (i.e., not augmented by the point at infinity) will be called the **finite complex plane**. Observe that if we are to allow the values $z = \infty$ and $w = \infty$, we need to

[*] Cleared of fractions, (1) becomes $czw + dw - az = b$, which is linear in z and linear in w; hence, it is *bilinear* in z and w.

decide how to interpret $(az+b)/(cz+d)$, say, for $z=\infty$. We define the latter as a/c if $c\neq 0$ and as ∞ if $c=0$.

Besides being conformal, the bilinear transformation is also *one-to-one* on the extended z and w planes, for (1) gives a unique w for each z, and the inverse mapping

$$z=-\frac{dw-b}{cw-a} \tag{3}$$

[obtained by solving (1) for z] gives a unique z for each w.

We will establish an important property of the bilinear transformation in a moment. To do so, we begin with the special case $w(z)=1/z$ and seek to determine the image (i.e., in the w plane) of any circle in the z plane. If the circle is centered at (a,b) and has radius c, its equation is

$$(x-a)^2+(y-b)^2=c^2 \tag{4}$$

or

$$x^2+y^2-2ax-2by=c^2-a^2-b^2. \tag{5}$$

To express the latter in terms of z, and hence w, we use the relations $x=(z+\overline{z})/2$ and $y=(z-\overline{z})/2i$. Then (5) becomes

$$z\overline{z}-Az-\overline{A}\overline{z}=B, \tag{6}$$

where $A=a-ib$ and $B=c^2-a^2-b^2$. But $z=1/w$ and $\overline{z}=1/\overline{w}$ so (6) gives

$$\frac{1}{w\overline{w}}-\frac{A}{w}-\frac{\overline{A}}{\overline{w}}=B \tag{7}$$

as governing the image curve.

Consider separately the cases $B\neq 0$ and $B=0$. If $B\neq 0$, then we can divide both sides of (7) by B. Doing so, and multiplying by $w\overline{w}$ as well, gives

$$w\overline{w}+\frac{\overline{A}}{B}w+\frac{A}{B}\overline{w}=\frac{1}{B}. \tag{8}$$

Since B is real, the coefficients of w and $\overline{w}$ are complex conjugates. Thus, (8) is of the same form as (6) so (8) represents a circle in the w plane.

Next, consider the case where $B=0$. Then (7) becomes

$$1-A\overline{w}-\overline{A}w=0 \tag{9}$$

or, with $A=a-ib$ and $w=u+iv$,

$$1-2au+2bv=0, \tag{10}$$

which is the equation of a straight line.

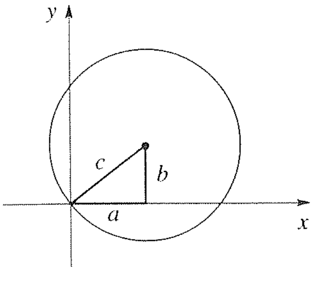


Figure 1. The case $B=0$.

We see from Fig. 1 that the geometrical significance of $B=0$ is that the circle in the z plane passes through the origin, because B is $c^2-a^2-b^2$. It makes sense that the image of a circle through $z = 0$ is a straight line because the mapping $w = 1/z$ sends $z = 0$ to $w = \infty$, and the only circle that reaches $w = \infty$ is a straight line – that is, a circle with infinite radius.

Similarly, we see that a straight line L in the z plane is mapped to a circle if L does not pass through $z = 0$, and to another straight line if L does pass through $z = 0$.

The upshot is that $w(z) = 1/z$ *sends circles into circles, with a straight line considered as a circle of infinite radius.* That result is not at all surprising if the original curve is a circle centered at $z = 0$ since if $|z| = R$ then $|w| = |1/z| = 1/R$, so the image is a circle of radius $1/R$ centered at $w = 0$. But it is remarkable that all circles are sent into circles, even if they are not centered at $z = 0$!

Now we turn from the special case $w(z) = 1/z$ to the general bilinear transformation (1). If we call

$$w = Az, \tag{11a}$$

$$w = z + B, \tag{11b}$$

$$w = \frac{1}{z} \tag{11c}$$

scaling and rotation, translation, and *inversion* transformations, respectively, then (1) amounts to a sequence, or composition, of transformations of these three types. For if $c \neq 0$, then it is readily verified (by successive substitution) that

$$\begin{aligned} w_1 &= cz, \qquad w_2 = w_1 + d, \qquad w_3 = \frac{1}{w_2}, \\ w_4 &= \frac{bc - ad}{c} w_3, \qquad w = w_4 + \frac{a}{c} \end{aligned} \tag{12}$$

takes us from z to w_1 to w_2 to w_3 to w_4 to w by a finite sequence of scaling and rotation, translation, and inversion transformations. Similarly if $c = 0$, for then

$$w_1 = \frac{a}{d} z, \qquad w = w_1 + \frac{b}{d}. \tag{13}$$

We call (11a) a scaling and rotation transformation because each z vector is scaled (by the modulus of the complex number A) and rotated (by the argument of A). It follows that the image region D' in the w plane will be a scaled and rotated version of the original region D in the z plane as illustrated in Fig. 2a.

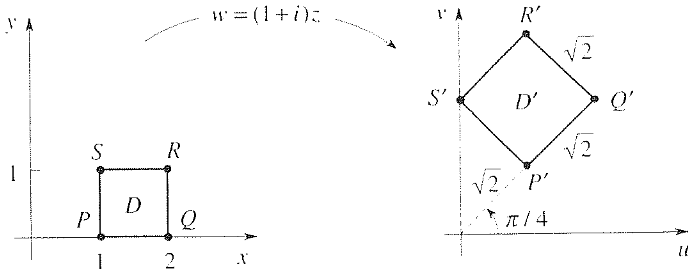


Figure 2. Scaling and rotation $w = Az$; $A = 1 + i$.

where $A = 1 + i$ and D is taken to be a unit square.

We call (11b) a translation transformation because each z point, and hence any given shape D, gets translated rightward by $\operatorname{Re} B$ and upward by $\operatorname{Im} B$. With $B = 1+2i$, for instance, the mapping $w = z+B = z+(1+2i)$ sends the region D in Fig. 2 into a square with corners at $2+2i$, $3+2i$, $3+3i$, and $2+3i$, respectively. And we've already shown that the inversion (11c) sends circles into circles. It should be clear that (11a) and (11b) do too, and since (1) is a composition of these three kinds of transformation (1) sends circles into circles, with straight lines as special cases of circles. That result is the most important property of the bilinear transformation.

THEOREM 22.3.1 *Circles Into Circles*
The bilinear transformation (1) sends every circle or straight line into either a circle or a straight line.

EXAMPLE 1. *Steady-State Temperature.* Find the steady-state temperature distribution $\psi(x,y)$ in the cut off disk shown in Fig. 3, subject to the boundary conditions shown there.

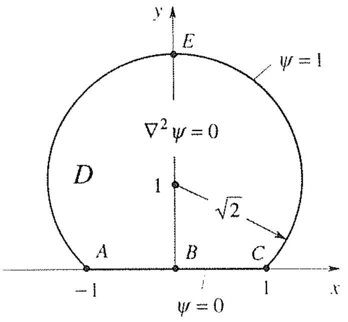


Figure 3. Steady-state temperature problem.

For the problem to be solvable by separation of variables we need the boundary of D to be comprised of constant-coordinate curves. Cartesian coordinates don't work because of the circular part of the boundary. Polar coordinates don't work either for if we locate the origin at the center of the circle, then the straight edge is neither an $r =$ constant curve nor a $\theta =$ constant curve. And if we locate the origin as in Fig. 3 then although BC and AB are constant θ curves ($\theta = 0$ and $\theta = \pi$, respectively), the circular part AEC is not an $r =$ constant curve.

Thus, let us try solving by means of conformal mapping. To do so, we need to find a conformal map $w(z)$ that sends the problem shown in Fig. 3 into one that is simpler. There could be many such maps, we simply need one.

Since ABC is straight and AEC is circular, the bilinear transformation warrants consideration. But how do we "design" that transformation – that is, how do we choose the parameters a, b, c, d in (1)? To begin, let us simply try the mapping $w = 1/z$ and see what happens. Since AEC is a circular arc, we know that its image will be too. So, to find its image, it suffices to find the image of the three points A, E, C, and to fit a circular arc through them as shown in Fig. 4; three points suffice because three noncollinear points uniquely determine a circle. [Note that the image of E is $1/(1+\sqrt{2})i = -(\sqrt{2}-1)i$.] The segment ABC is also a circular arc (of infinite radius), but we need to be careful because B is the pole of the transformation and gets sent to infinity. Thus, we need to treat AB and BC separately. Consider AB. Since B goes to infinity, the image of AB is a straight line that extends from the image of A (namely, A') to infinity. To determine that line we need one more point so consider $z = -1/2$: A goes to -1 and $z = -1/2$ goes to $w = -2$ so the image of AB is as shown in Fig. 4, extending from $u = -1$ to $u = -\infty$ on the u axis. Similarly for BC: B goes to infinity, C goes to 1, and (since one more point is needed) $z = 1/2$ goes to $w = 2$ so the image of BC is as shown in Fig. 4, extending from $u = 1$ to $u = \infty$ on the u axis.

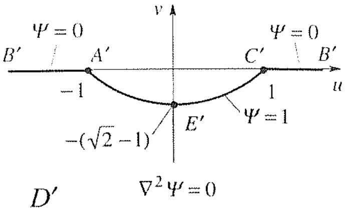


Figure 4. The mapped problem, using $w = 1/z$.

Do not be concerned that B' shows up in two different places in Fig. 4 we simply call B' "the point at infinity." It is approached from different directions, according to how B is approached in the z plane.

Our conclusion is that the mapping $1/z$ is of no help because the problem in Fig. 4 is no easier than the one in Fig. 3. Trying again, let us design the bilinear transformation so as to put the pole at A or C, say at A. If, for simplicity, we also send C to $w = 0$, then we have the mapping

$$w(z) = \frac{z-1}{z+1}. \tag{14}$$

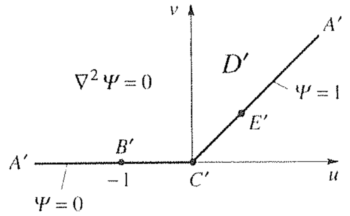


Figure 5. The mapped problem using $w = (z-1)/(z+1)$.

Then A goes to infinity, B goes to $w = -1$, C goes to $w = 0$, and E goes to $w = \frac{1+\sqrt{2}}{2+\sqrt{2}}(1+i)$, which has an argument of $\pi/4$. Thus, under the mapping (14) the Ψ problem is as shown in Fig. 5. To solve, introduce polar coordinates ρ, ϕ with $u = \rho\cos\phi$, $v = \rho\sin\phi$. Then, expressing $\nabla^2\Psi = 0$ as $\Psi_{\rho\rho} + \frac{1}{\rho}\Psi_\rho + \frac{1}{\rho^2}\Psi_{\phi\phi} = 0$ and using separation of variables as in Section 20.3, we obtain

$$\Psi(\rho,\phi) = (A + B\ln\rho)(C + D\phi) + (E\rho^\kappa + F\rho^{-\kappa})(G\cos\kappa\phi + H\sin\kappa\phi). \tag{15}$$

To keep Ψ bounded as $\rho \to \infty$ and as $\rho \to 0$ (and hence ψ bounded as $z \to -1$ and $z \to 1$, respectively) set $B = E = F = 0$ so Ψ is of the form

$$\Psi(\rho,\phi) = C_1 + C_2\phi. \tag{16}$$

The boundary conditions $\Psi(\rho,\pi/4) = 1$ and $\Psi(\rho,\pi) = 0$ give C_1 and C_2 so

$$\Psi(\rho,\phi) = \frac{4}{3}\left(1 - \frac{\phi}{\pi}\right).$$

To complete the solution we need to express ρ and ϕ [actually, just ϕ because there are no ρ's in (16)] in terms of x and y. Equation (14) gives

$$u + iv = \frac{(x-1)+iy}{(x+1)+iy}\,\frac{(x+1)-iy}{(x+1)-iy} = \frac{(x^2+y^2-1)+2iy}{(x+1)^2+y^2} \tag{17}$$

so

$$u = \frac{x^2+y^2-1}{(x+1)^2+y^2}, \qquad v = \frac{2y}{(x+1)^2+y^2}. \tag{18}$$

Finally, since $\phi = \tan^{-1}(v/u)$, the desired solution is

$$\psi(x,y) = \frac{4}{3}\left[1 - \frac{1}{\pi}\tan^{-1}\left(\frac{2y}{x^2+y^2-1}\right)\right], \tag{19}$$

where the $\tan^{-1}$ lies between $\pi/4$ and π. For instance, at $x = y = 1$ we obtain $\psi(1,1) = 0.863$ which, from Fig. 3, seems reasonable.

COMMENT. Having found the boundary of D', in Fig. 5, to be the bent line $A'B'C''E'A'$, how did we know that the region D' was above it rather than below it? To answer that question it suffices to check a single point in the interior of D, say $z = i$. That point is mapped by (14), into $w = i$, so the region of D' is above $A'B'C'E'A'$ as shown. ▮

EXAMPLE 2. Solve the potential problem shown in Fig. 6a. Since the boundary of D

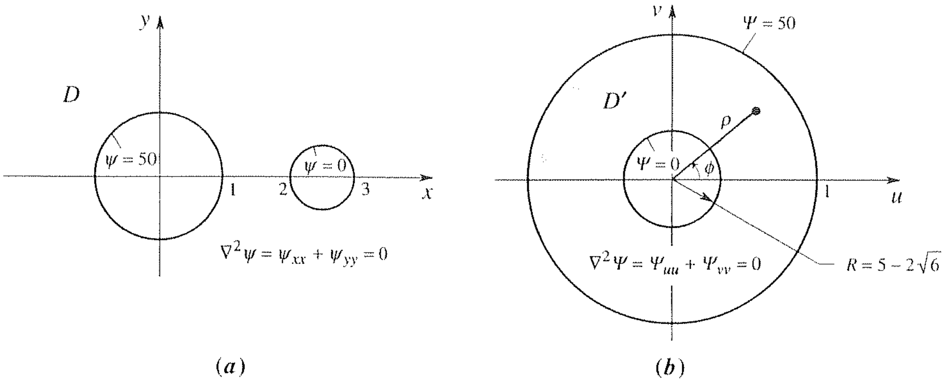


Figure 6. Region with two cutouts.

is comprised of circles, we consider the bilinear transformation once again. How are we to determine the parameters a, b, c, d in (1) so that the region D is sent into something simple, such as a region bounded by *concentric* circles? Since three noncollinear points determine a circle, we could seek a, b, c, d so as to map three selected points on each of the two circles in the z plane into three selected points on each of two concentric circles in the w plane. That procedure would give six equations – actually twelve because in each equation we would equate real and imaginary parts on the left- and right-hand sides. How many parameters do we have? Each of a, b, c, d has a real and an imaginary part so we have eight parameters. But we can divide both numerator and denominator in (1) by one of them, to normalize, so there are really seven "design parameters," and twelve linear algebraic equations for them. In general, such a system is overconstrained and there is no solution. However, the number of free parameters can be increased if we use the fact that we don't need to choose three specific points on each of the circles in the z plane; *any* three points on each of those circles will do. Similarly for the points chosen on the image circles. Further, we don't care what the radii of the image circles are, so we can let them be free parameters.

The upshot is that it is possible to map any two distinct circles in the z plane into two concentric circles in the w plane, but design of such a map is tedious and best avoided. Thus, it is much more convenient to seek a suitable mapping in a table of mappings. A short table is given in Appendix F, and the desired mapping is given by entry 4,

$$w = \frac{z - a}{az - 1}, \tag{20}$$

where a is defined in terms of the x-axis intercepts of the right-hand circle, $x_1 = 3$ and $x_2 = 2$. With these values $a = (7 + 2\sqrt{6})/5$, and the image is as shown in Fig. 6b.

To solve the problem on Ψ, introduce polar coordinates ρ, ϕ ($u = \rho\cos\phi, v = \rho\sin\phi$), so that

$$\nabla^2\Psi = \Psi_{\rho\rho} + \frac{1}{\rho}\Psi_\rho + \frac{1}{\rho^2}\Psi_{\phi\phi} = 0. \tag{21}$$

Because the annulus is axisymmetric and the boundary conditions do not vary with ϕ, we can find a solution $\Psi(\rho)$ that varies only with ρ. Then the $\Psi_{\phi\phi}$ term in (21) drops out and (21) simplifies to the ordinary differential equation

$$\frac{d^2}{d\rho^2}\Psi + \frac{1}{\rho}\frac{d}{d\rho}\Psi = 0, \tag{22}$$

with general solution $\Psi(\rho) = A + B\ln\rho$. [Or, we could obtain the latter by setting the coefficients of all ϕ-dependent terms in (15) equal to zero.] Using the boundary conditions $\Psi(R) = 0$ and $\Psi(1) = 50$ to evaluate A and B gives

$$\Psi(\rho) = 50\left(1 - \frac{\ln\rho}{\ln R}\right). \tag{23}$$

Finally, obtaining ρ in terms of x and y (which step is left for the exercises) gives

$$\begin{aligned}\psi(x,y) &= \Psi(\rho(x,y)) \\ &= \frac{50}{\ln R}\left[\ln R - \frac{1}{2}\ln\frac{(x-a)^2+y^2}{(ax-1)^2+a^2y^2}\right]\end{aligned} \tag{24}$$

as the desired solution. Representative equipotentials are plotted in Fig. 7. ▮

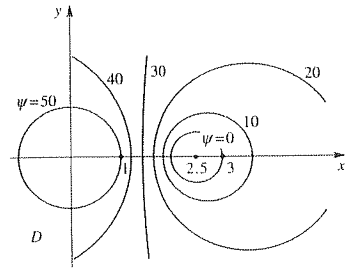


Figure 7. Representative equipotentials, from (24).

Closure. In this section we study only the bilinear transformation (1), the most important property of which is that it sends circles into circles (with straight lines as special cases of circles). That property makes the bilinear transformation useful for problems in which the domain D is bounded by circles or circular arcs. That is not to say that we cannot use a bilinear transformation for other types of regions, but it is unlikely that a simplification of the domain can thereby be accomplished. Nor is it true that every domain D that is bounded by circular arcs can be handled with a bilinear transformation since the seven design parameters might not be enough. For instance, a region with *three* or more circular cutouts is simply more than can be handled.

In any given application, be sure to verify that the mapping function $w(z)$ is analytic and that $w'(z) \neq 0$ everywhere within the interior of D. For the mapping given by (20), for instance, $w(z)$ is analytic for all z except $z = 1/a \approx 0.42$, which falls within one of the cutouts rather than in D, and $w'(z) = (a^2-1)/(az-1)^2 \neq 0$ everywhere.

Finally, we suggest that simplified problems to aim at, in the w plane, are ones in which D' is either an annulus bounded by concentric circles (Fig. 6b) or an infinite wedge (Fig. 5) or strip, with constant boundary conditions.

EXERCISES 22.3

1. We stated that (8) represents a circle in the w plane. Give the equation of that circle in the form $(u-\alpha)^2 + (v-\beta)^2 = \gamma^2$, and give α, β, γ in terms of the quantities a, b, c that appear in (4).

2. Following essentially the same steps as we used in equations (4)–(8) to show that $w(z) = 1/z$ sends circles into circles, show that the following transformation also sends circles into circles.

(a) $w(z) = Pz$, where P is nonzero and, in general, complex.
(b) $w(z) = z + Q$, where Q is nonzero and, in general, complex.

3. (a) Derive the expression

$$\rho(x,y) = \ln\sqrt{\frac{(x-a)^2+y^2}{(ax-1)^2+a^2y^2}}$$

that was used in (24).

(b) Use computer software, such as the *Maple* implicitplot command, to obtain the isotherms shown in Fig. 7.

4. (*Composition of two bilinear transformations*) Show that the composition of two bilinear transformations is also a bilinear transformation. That is, show that if $f(z)$ and $g(z)$ are bilinear transformations so is $f(g(z))$.

5. (*Fixed points*) A **fixed point** of a mapping $w = f(z)$ is a point z_0 that is mapped into itself. That is, z_0 is a fixed point of $w = f(z)$ if $f(z_0) = z_0$. Prove that a *bilinear transformation* $w = (az+b)/(cz+d)$ *has at most two fixed points, unless it is simply the identity transformation* $w = z$.

6. (*Three points into three points*) (a) Prove that *any three given distinct points* z_1, z_2, z_3 *in the extended* z *plane can be mapped into any three given distinct points* w_1, w_2, w_3, *respectively, in the extended* w *plane by a bilinear transformation* $w = f(z)$ *which is given, implicitly, by the relation*

$$\frac{w - w_1}{w - w_3}\,\frac{w_2 - w_3}{w_2 - w_1} = \frac{z - z_1}{z - z_3}\,\frac{z_2 - z_3}{z_2 - z_1}. \tag{6.1}$$

HINT: Show first that the transformation $w = f(z)$ implied by (6.1) is indeed bilinear. Then verify that (6.1) guarantees that $f(z_1) = w_1$, $f(z_2) = w_2$, and $f(z_3) = w_3$.
(b) Prove that the bilinear transformation $w = f(z)$ in part (a) is unique. HINT: Suppose that $w = g(z)$ is another bilinear transformation which maps the three given z points into the three given w points. With the help of the results stated in Exercises 4 and 5, show that the composite transformation $g^{-1}(f(z))$ has three fixed points so it must be the identity transformation.

7. Verify that if a is a complex number with $|a| < 1$, then $w = (z - a)/(1 - \overline{a}z)$ maps $|z| < 1$ onto $|w| < 1$, with a being sent to the origin.

8. (*On which side of the boundary curve is* D'?) First, review the COMMENT at the end of Example 1. Rather than rely on a pointwise check, one can determine whether D' is on one side of the mapped boundary curve or the other, as follows. *Let three points on the boundary curve of* D *be labeled consecutively as* A, B, C, *and imagine walking along that curve from* A *to* B *to* C. *If the region is on our left (right), then likewise the region* D' *is on our left (right) when we walk along the corresponding curve in the* w *plane, from* A' *to* B' *to* C'. For instance, when we walk along ABC in Fig. 3 the region D is on our left, so when we walk along $A'B'C'$ in Fig. 5 the region D' must, likewise, be on our left. The problem is this. Show that the truth of the italicized claim, above, follows from the conformality of the mapping.

9. (a) Use (6.1) in Exercise 6 to find two distinct bilinear transformations that map $x < 0$ onto $|w| < 2$, and express them in the form $w = (az+b)/(cz+d)$. HINT: Use the idea in Exercise 8 to be sure that D' is $|w| < 2$ rather than $|w| > 2$.
(b) The same as (a), but this time map $x < 0$ onto $|w| > 2$.
(c) The same as (a), but this time map $x+y < 1$ onto $u-v > 3$.
(d) The same as (a), but this time map $x+y > 4$ onto $|w| < 1$.

10. Determine the image of the given region under the mapping $w = (z+1)/(z-i)$. Label the various key points in the z and w planes.

(a) $y < 0$ (b) $0 < y < 1$
(c) $y > 1$ (d) $x > 0$
(e) the wedge $0 < \arg z < \pi/4$
(f) the quadrant $0 < \arg z < \pi/2$
(g) the annulus $1 < |z| < 2$

11. Determine the image of the region $0 < x < \infty$, $0 < y < \infty$ under the given mapping. Label any key points.

(a) $w = 1/(z-i)$ (b) $w = 1/(z+i)$
(c) $w = z/(z-1)$ (d) $w = 2 - i/z$
(e) $w = (z-1)/(z+2)$ (f) $w = (z-3)/z$

12. Putting the pole of the transformation at A, we used (14) to solve the problem shown in Fig. 3. Putting the pole at C instead, use

$$w(z) = \frac{z+1}{z-1},$$

and show that your final result agrees with that given by (19).

13. Use computer software and equation (19) to plot the isothermal curves $u = 0, 0.2, 0.4, 0.6, 0.8$, and 1. Using *Maple*, for instance, this can be done using the implicitplot command.

14. Solve the given Dirichlet problem with the help of a bilinear transformation. Obtain that transformation yourself, or use the table in Appendix F. In each case the PDE is $\nabla^2\psi = \psi_{xx} + \psi_{yy} = 0$.

(a) Let D be the infinite plane, with a cutout defined by $x^2 + (y-3)^2 = 25$ for $y \geq 0$ and by $x^2 + (y+3)^2 = 25$ for $y \leq 0$. Let $\psi = 100$ on the former, and $\psi = 0$ on the latter. HINT: D can be mapped onto an infinite wedge.
(b) Let D be the region $x < 2$ with the disk $x^2 + y^2 \leq 1$ cut out. Let $\psi = 30$ on $x = 2$ and $\psi = 20$ on $x^2 + y^2 = 1$. HINT: The desired mapping can be obtained as a limiting case of the one given by item 4 of Appendix F.
(c) Let D be the disk $x^2+y^2 < 9$, with the disk $(x-\frac{3}{2})^2+y^2 \leq \frac{9}{4}$ cut out. Let $\psi = 1$ on $x^2 + y^2 = 9$ and $\psi = 0$ on $(x - \frac{3}{2})^2 + y^2 = \frac{9}{4}$. HINT: Locate the pole of the transformation at a suitable point on the boundary of D.

(d) Let D be the "crescent" bounded above by the circle passing from -3 to $9i$ to 3, and below by the circle passing from -3 to $3i$ to 3. Let $\psi = 1$ on the upper arc and $\psi = 0$ on the lower arc. In particular, evaluate $\psi(0, 4)$. Also, explain why D cannot be mapped, by a bilinear transformation, onto an infinite strip or the region between concentric circles. HINT: Map D onto an infinite wedge.

(e) Let D be the disk $|z| < 1$ with the disk $\left|z - \frac{1}{4}\right| \le \frac{1}{4}$ cut out. Let $\psi = 50$ on $|z| = 1$ and $\psi = 0$ on $\left|z - \frac{1}{4}\right| = \frac{1}{4}$.

22.4 Additional Mappings and Applications

The bilinear transformation, discussed in Section 22.3, is nontypical in the sense that it is the only transformation that is one-to-one. That is, $w = (az + b)/(cz + d)$ gives a unique w for each z (except for $z = -d/c$), and solving for z gives the unique value $z = (-dw + b)/(cw - a)$ for each w (except for $w = a/c$).

Thus, with any other mapping we will inevitably need to deal with multivaluedness. For brevity, we will consider just two representative examples, with other mappings and applications reserved for the exercises.

EXAMPLE 1. Solve for the potential $\psi(x, y)$ within the curved strip shown in Fig. 1, subject to the boundary conditions shown there, together with a condition that ψ be bounded on D.

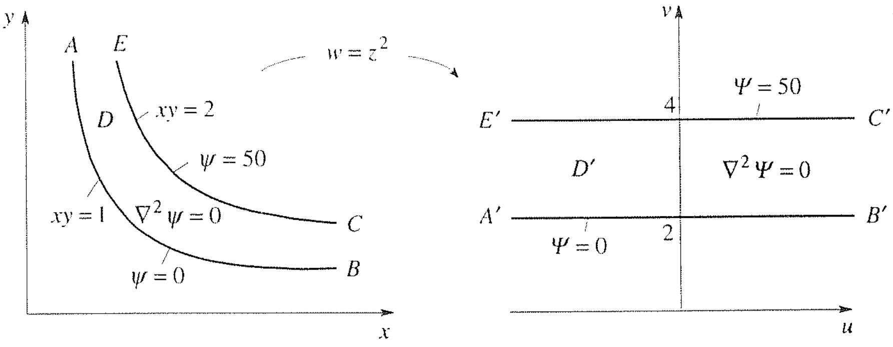


Figure 1. Potential in a curved strip.

A suitable mapping for the curved strip is not in the short table provided in Appendix F,* but is given by

$$w = z^2 = (x^2 - y^2) + i\,2xy \tag{1}$$

because we see from

$$u = x^2 - y^2, \qquad v = 2xy \tag{2}$$

that AB and EC map into the straight lines $v = 2$ and $v = 4$, over $-\infty < u < \infty$. Thus, the problem on $\Psi(u, v)$ shown in Fig. 1 is simple: we can seek Ψ as a function of v only so

*For a much more extensive table, see for example H. Kober, *Dictionary of Conformal Representation* (New York: Dover, 1952).

$\nabla^2\Psi = \Psi_{uu} + \Psi_{vv} = 0$ reduces to $d^2\Psi/dv^2 = 0$. The general solution is $\Psi = Av + B$, and the boundary conditions give

$$\Psi = 25(v-2). \tag{3}$$

Finally, recalling from (2) that $v = 2xy$, we have the desired solution

$$\psi(x,y) = 50(xy-1). \tag{4}$$

COMMENT. We mentioned above, that multi-valuedness was inevitable. Indeed, although $w = z^2$ is single-valued, the inverse transformation $z = \sqrt{w}$ is double-valued. That statement might be surprising inasmuch as when we returned from the w plane [equation (3)] to the z plane [equation (4)] we did not seem to have to make a choice between two values. Actually we did, but it might have gone unnoticed. Namely, the region D' maps back into D in the first quadrant but also into a similar curved strip between $xy = 1$ and $xy = 2$ in the third quadrant. Without fuss, we simply chose the one in the first quadrant. We could have introduced a branch cut for the $\sqrt{w}$ function, but it really wasn't necessary. (See Exercise 1.) ▮

EXAMPLE 2. *Electric Potential in Semi-Infinite Strip.* We wish to solve for the electric potential (i.e., the voltage) $\psi(x,y)$ within the semi-infinite strip shown in Fig. 2. The edges $x = 0$ and $x = 1$ are "grounded" ($\psi = 0$), the edge $y = 0$ is maintained at 100 volts, and we ask ψ to be bounded on D.

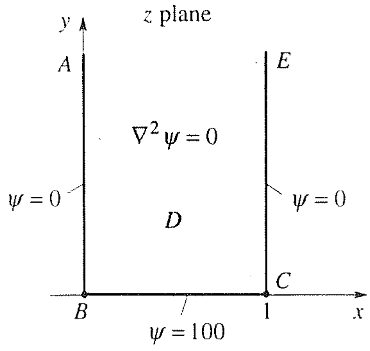


Figure 2. Electric potential in semi-infinite strip.

It's true that the boundary of D is comprised of circular arcs (namely, straight lines), but the bilinear transformation is of no help (Exercise 2). Turning to Appendix F for help, let us try the mapping $w = -\cos\pi z$ (entry 10). Then the mapped problem is as shown in Fig. 3. That problem is easier in the sense that it is the Dirichlet problem for the upper half plane, which has already been solved using a Fourier transform in Section 20.4. Using the solution given in that section, we have

$$\Psi(u,v) = \frac{v}{\pi}\int_{-\infty}^{\infty}\frac{\Psi(u',v)du'}{(u'-u)^2+v^2} = \frac{100v}{\pi}\int_{-1}^{1}\frac{du'}{(u'-u)^2+v^2}$$

$$= \frac{100v}{\pi}\int_{-1-u}^{1-u}\frac{d\xi}{\xi^2+v^2} = \frac{100}{\pi}\left[\tan^{-1}\left(\frac{1-u}{v}\right) - \tan^{-1}\left(\frac{-1-u}{v}\right)\right], \tag{5}$$

where $-\pi/2 < \tan^{-1}(\,) < \pi/2$ for each of the $\tan^{-1}(\,)$'s. [As a check, observe that for $u = 0$ and $v \to 0$ equation (1) gives $\Psi \to \frac{100}{\pi}(\frac{\pi}{2}+\frac{\pi}{2}) = 100$, for $u > 1$ and $v \to 0$ equation (5) gives $\Psi \to 0$, and for $u < -1$ and $v \to 0$ equation (5) gives $\Psi \to 0$, all of which are correct.]

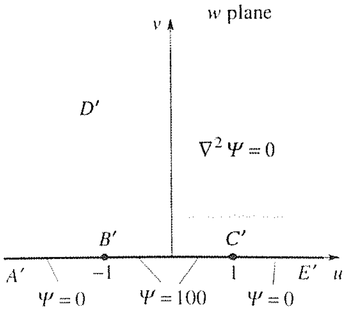


Figure 3. The mapped problem; $w = -\cos\pi z$.

Next, $w = -\cos\pi z = -\cos\pi(x+iy) = -\cos\pi x\cosh\pi y + i\sin\pi x\sinh\pi y$ gives

$$u = -\cos\pi x\cosh\pi y, \qquad v = \sin\pi x\sinh\pi y, \tag{6}$$

and putting these expressions into (5) gives the solution

$$\psi(x,y) = \frac{100}{\pi}\left[\tan^{-1}\left(\frac{\cos\pi x\cosh\pi y + 1}{\sin\pi x\sinh\pi y}\right) - \tan^{-1}\left(\frac{\cos\pi x\cosh\pi y - 1}{\sin\pi x\sinh\pi y}\right)\right], \tag{7}$$

where we recall that $-\pi/2 < \tan^{-1}(\,) < \pi/2$ for each $\tan^{-1}(\,)$.

If we do not notice that the w plane problem can be solved by the result given in Section 20.4, we can make one more transformation, from the w plane to a ζ plane (where $\zeta = \xi + i\eta$) using entry 8 of Appendix F. Denoting $\Psi(u(\xi,\eta), v(\xi,\eta)) \equiv \tilde{\Psi}(\xi,\eta)$, the result of this second mapping is indicated in Fig. 4. Finally, the problem in the ζ plane is simple (since $\tilde{\Psi}$ varies with η but not ξ) and gives the solution

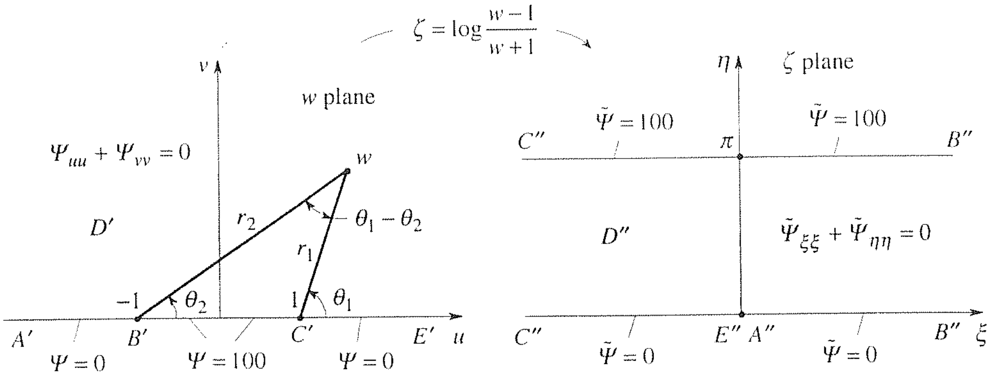


Figure 4. One more mapping.

$$\tilde{\Psi}(\xi,\eta) = \frac{100}{\pi}\eta. \tag{8}$$

Retracing our steps, we need to express η in terms of u and v, and then u and v in terms of x and y. First, we express $w - 1 = r_1 e^{i\theta_1}$ and $w + 1 = r_2 e^{i\theta_2}$ (Fig. 4) so that

$$\zeta = \xi + i\eta = \log\left(\frac{r_1}{r_2}e^{i(\theta_1-\theta_2)}\right) = \ln\frac{r_1}{r_2} + i(\theta_1 - \theta_2). \tag{9}$$

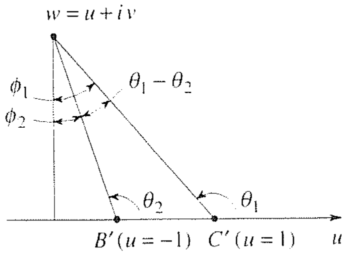


Figure 5. The angle $\theta_1 - \theta_2$.

It follows from (9) that $\eta = \theta_1 - \theta_2$ so

$$\Psi(u,v) = \frac{100}{\pi}(\theta_1 - \theta_2). \tag{10}$$

It is convenient to use the fact that the angle $\theta_1 - \theta_2$ in Fig. 4 is equivalent to $\phi_1 - \phi_2$ (Fig. 5) so

$$\Psi(u,v) = \frac{100}{\pi}(\phi_1 - \phi_2) = \frac{100}{\pi}\left[\tan^{-1}\left(\frac{1-u}{v}\right) - \tan^{-1}\left(\frac{-1-u}{v}\right)\right], \tag{11}$$

where $-\pi/2 < \tan^{-1}(\,) < \pi/2$ for each $\tan^{-1}(\,)$, and we see that this result is identical to (5).

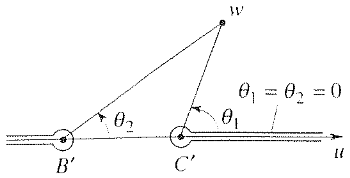


Figure 6. Branch cut chosen for $\log[(w-1)/(w+1)]$.

COMMENT 1. We didn't say so explicitly, but the branch cut that we used to make $\zeta = \log\frac{w-1}{w+1} = \log(w-1) - \log(w+1)$ single-valued is as shown in Fig. 6. Other choices could have been made, but the final solution would have been unaffected.

COMMENT 2. Observe that we mapped from the z plane to the w plane using $w = -\cos\pi z$, and then from the w plane to the ζ plane using $\zeta = \log\frac{w-1}{w+1}$. The two steps could have been combined into one as

$$\zeta = \log\left(\frac{\cos \pi z + 1}{\cos \pi z - 1}\right), \tag{12}$$

but it is much easier to proceed by a sequence of small steps – both in this problem and as a general rule.

COMMENT 3. In this example, the original semi-infinite strip problem could have been solved by separation of variables, which method gives (Exercise 3)

$$\psi(x, y) = \frac{400}{\pi} \sum_{n=1,3,\dots}^{\infty} \frac{\sin n\pi x}{n} e^{-n\pi y}, \tag{13}$$

but surely the conformal mapping result (7) is nicer since it is in closed form, whereas the separation of variable result (13) is in the form of an infinite series. We should be able to obtain (13) from (7) by expressing

$$\cosh \pi y = \frac{e^{\pi y} + e^{-\pi y}}{2} = \left(\frac{1}{p} + p\right)\Big/2 = (1 + p^2)/2p,$$
$$\sinh \pi y = \frac{e^{\pi y} - e^{-\pi y}}{2} = \left(\frac{1}{p} - p\right)\Big/2 = (1 - p^2)/2p,$$

where $p = e^{-\pi y}$, and doing a Taylor expansion of the right-hand side of (7) in powers of p, but we will not pursue that point. ▮

Closure. Important points made in this section are as follows. First, in selecting a mapping, for a given application, it is reasonable to rely on conformal mapping tables. Second, one often employs not a single mapping but a sequence of mappings, as we illustrate in Example 2 and as are illustrated in the exercises of Section 22.6. Third, if we use a mapping other than the bilinear transformation, then it will inevitably involve multi-valuedness – that is, it will not be one-to-one. Finally, in seeking the image of a circular arc under a bilinear transformation, in Section 22.3, it sufficed to map only three points (two, in fact, if we know the image to be a straight line) because three noncollinear points uniquely determine the circular image curve. In general, however, the image of a given curve, under a given mapping, is not a circle, and three points by no means suffice. Rather, seek the image of the entire curve, not of individual points. For instance, if a given curve C in the z plane is given parametrically by $x = x(\tau)$ and $y = y(\tau)$, for $a < \tau < b$, then the image curve is given parametrically by $u = u(x(\tau), y(\tau))$ and $v = v(x(\tau), y(\tau))$, for $a < \tau < b$. As a simple illustration, let us take another look at finding the image of the curve $xy = 2$, in Example 1 (Fig. 1). If we parametrize that curve by

$$x = \tau, \quad y = 2/\tau, \qquad (0 < \tau < \infty)$$

then (2) gives

$$u = \tau^2 - 4/\tau^2, \quad v = 4, \qquad (0 < \tau < \infty)$$

which is the horizontal line $v = 4$, for $-\infty < u < \infty$.

EXERCISES 22.4

1. Show a branch cut for $z = \sqrt{w}$ that sends the infinite strip D' (Fig. 1) back into the curved strip $1 < xy < 2$ in the first quadrant, rather than in the third quadrant.

2. Show that the region D in Fig. 2 cannot be mapped into a circular annulus or an infinite strip by any bilinear transformation. HINT: Note that D has two corners on its boundary.

3. Derive the separation of variables solution (13) to the problem on $\psi(x, y)$ in Example 2.

4. In each case, solve $\nabla^2\psi = 0$ for $\psi(x, y)$ for the specified region D and the specified boundary conditions, together with the condition that ψ be bounded on D. If useful, you may use the known solution of the Dirichlet problem for the half plane [(12) in Section 20.4], as we did in Example 2.

(a) Let D be the strip $0 < y < \pi$, $-\infty < x < \infty$. Let $\psi = 0$ everywhere on the boundary except on $y = 0$ $(-\infty < x < 0)$, where $\psi = 20$. Also, evaluate $\psi(0, \pi/2)$. HINT: Use entry 6 in Appendix F.
(b) The same as part (a) but with $0 < y < 5$ instead of $0 < y < \pi$. HINT: Modify the mapping given in entry 6 slightly, so that D' is once again the upper half plane.
(c) Let D be the region $0 < x < \infty$, $0 < y < \infty$. Let $\psi = 0$ everywhere on the boundary except on $x = 0$ $(2 < y < \infty)$, where $\psi = 35$. Also, evaluate $\psi(2, 2)$. HINT: Use entry 5 in Appendix F.
(d) The same as part (c), but with different boundary conditions. This time, let $\psi = 0$ everywhere on the boundary except on $x = 0$ $(0 < y < 2)$, where $\psi = 35$.
(e) Let D be the 45° wedge between the positive x axis and the line $y = x$. Let $\psi = 0$ everywhere on the boundary except on the x axis from $x = 5$ to $x = \infty$, where $\psi = 200$. Also evaluate $\psi(5, 1)$. HINT: Use entry 5 in Appendix F.
(f) The same as part (e) but with different boundary conditions. This time let $\psi = 10$ everywhere on the boundary except on the x axis from $x = 0$ to $x = 10$, where $\psi = 0$.
(g) Let D be the semi-infinite strip $0 < x < 1$, $0 < y < \infty$. Let $\psi = 0$ on $x = 0$ and on $y = 0$, and let $\psi = 400$ on $x = 1$. Also, evaluate $\psi(0.5, 0.1)$. HINT: Use entry 10 in Appendix F.
(h) Let D be the semi-infinite strip $0 < x < 4$, $0 < y < \infty$. Let $\psi = 0$ on $y = 0$, and let $\psi = 200$ on $x = 0$ and $x = 4$. Also, evaluate $\psi(2, 2)$, $\psi(2, 4)$, $\psi(2, 6)$, $\psi(2, 8)$, $\psi(2, 10)$, and $\psi(2, 20)$. HINT: You can use entry 10 in Appendix F, but will need to modify it slightly.
(i) Let D be the semi-infinite strip $-2 < x < 2$, $0 < y < \infty$. Let $\psi = 0$ on $x = -2$, and on $y = 0$ $(-2 < x < 0)$, and let $\psi = 50$ on $x = 2$ and $y = 0$ $(0 < x < 2)$. Also, evaluate $\psi(0, 1)$ and $\psi(0, 30)$. HINT: With a slight modification you can use entry 12 in Appendix F.
(j) Let D be the region in the right half plane, between the lines $y = \pm x$ and the curve $x^2 - y^2 = 4$. Let $\psi = 30$ on $y = \pm x$ and let $\psi = 20$ on $x^2 - y^2 = 4$. Also, evaluate $\psi(1, 0)$ and $\psi(0.6, -0.4)$. HINT: Use the mapping $w(z) = z^2$.

22.5 More General Boundary Conditions

Thus far, all of the boundary conditions considered have been constant and of Dirichlet type, $\psi =$ constant. In this section we show how to handle nonconstant boundary conditions, Dirichlet conditions (ψ given), and Neumann conditions ($\partial\psi/\partial n$ given) as well.

The fate of a **Dirichlet boundary condition**, under a mapping $w(z)$, is simple. Since

$$\boxed{\psi(x, y) = \psi(x(u, v), y(u, v)) \equiv \Psi(u, v),} \tag{1}$$

we see that if ψ takes on the value 26, say, at a point P on the boundary of D, then

Ψ takes on that same value at the point P' (the image of P) on the boundary D'.

EXAMPLE 1. *Nonconstant Dirichlet conditions.* Find the boundary conditions in the w plane for the problem displayed in Fig. 1. Here, D is the region between the hyperbola $xy = 3$ and the positive x and y axes, and the $w = z^2$ mapping gives the relations

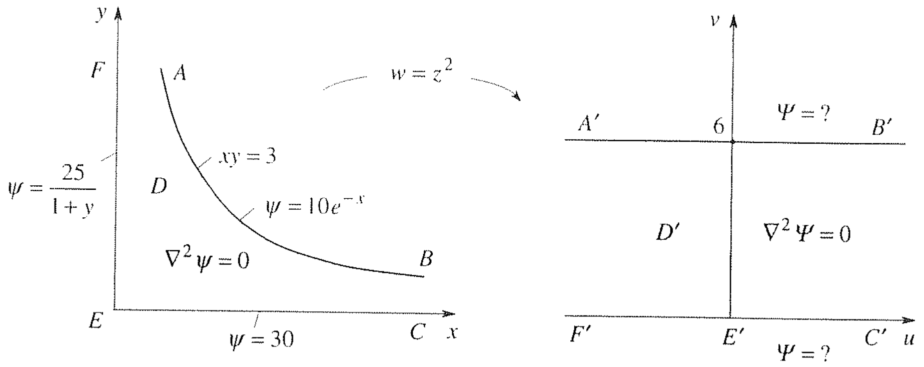


Figure 1. Nonconstant Dirichlet conditions.

$$u = x^2 - y^2, \qquad v = 2xy. \tag{2}$$

The $\psi = 30$ boundary condition on EC carries over to $\Psi = 30$ on its image $E'C'$. To determine the image of the $\Psi = 10e^{-x}$ condition on AB, we need to express x (on AB) in terms of u (on $A'B'$). To do so, put $y = 3/x$ into (2), obtaining

$$u = x^2 - \frac{9}{x^2}, \qquad v = 6. \tag{3}$$

Multiplying the first of these by x^2 gives the $x^4 - ux^2 - 9 = 0$, which is a quadratic equation in x^2, so that

$$x^2 = \frac{u \pm \sqrt{u^2 + 36}}{2}, \qquad x = \sqrt{\frac{u + \sqrt{u^2 + 36}}{2}}, \tag{4}$$

where, in the last step, we selected the + sign since x is to be real. Thus, the boundary condition on $A'B'$ is

$$\Psi(u, 6) = 10 \exp\left(-\sqrt{\frac{u + \sqrt{u^2 + 36}}{2}}\right). \qquad (-\infty < u < \infty) \tag{5}$$

On EF we have $x = 0$ so (2) gives $u = -y^2$ and $v = 0$. Thus, $y = \sqrt{-u}$ (remember that $u < 0$ on $F'E'$) and the boundary condition on $F'E'$ is

$$\Psi(u, 0) = \frac{25}{1 + \sqrt{-u}}. \qquad (-\infty < u < 0) \tag{6}$$

We will not attempt to solve the Ψ problem. Our purpose was simply to illustrate the transformation of the nonconstant Dirichlet boundary conditions. ▮

Next, we turn to **Neumann boundary conditions**, namely, where $\partial\psi/\partial n$ is prescribed on the boundary of D, $\partial\psi/\partial n$ being the directional derivative of ψ in the outward normal direction. How does $\partial\psi/\partial n$ transform under the mapping $w = f(z)$?

Let z be any point on the boundary C of D, and let z_0 lie within D on a straight line L which is normal to C at z (Fig. 2). The image L' of L will, in general, be curved, but conformality guarantees that it be normal to C' at w. Then, with $\Delta z = z - z_0$ and $\Delta w = w - w_0$ we have

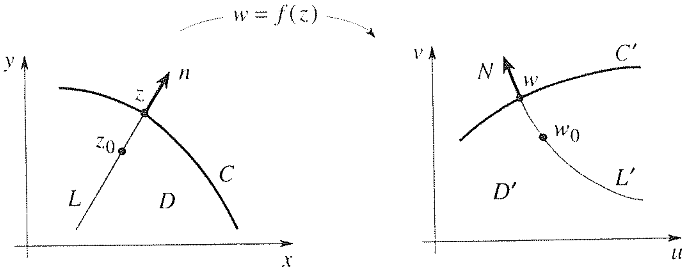


Figure 2. Transformation of Neumann condition.

$$\left.\frac{\partial\psi}{\partial n}\right|_z = \lim_{\Delta z\to 0} \frac{\psi|_z - \psi|_{z_0}}{|\Delta z|}. \tag{7}$$

But $\psi|_z = \Psi|_w$, $\psi|_{z_0} = \Psi|_{w_0}$, and $\lim_{\Delta z\to 0} \Delta w/\Delta z = f'(z)$ so

$$\left.\frac{\partial\psi}{\partial n}\right|_z = \lim_{\Delta z\to 0} \frac{\Psi|_w - \Psi|_{w_0}}{|\Delta w|}\frac{|\Delta w|}{|\Delta z|} = \left.\frac{\partial\Psi}{\partial N}\right|_w |f'(z)|, \tag{8}$$

where $\partial\Psi/\partial N$ denotes the directional derivative of Ψ in the outward normal direction. In the final equality in (8) we used the fact that $\Delta z \to 0$ implies that $\Delta w \to 0$ for if $w = f(z)$ is a differentiable function of z, then it must surely be a continuous function of z. We conclude that Neumann boundary conditions transform, under a conformal map, according to

$$\boxed{\frac{\partial\Psi}{\partial N} = \frac{1}{|f'(z)|}\frac{\partial\psi}{\partial n},} \tag{9}$$

where $f'(z) \neq 0$ by the assumed conformality. As an important special case, we see from (9) that if $\partial\psi/\partial n = 0$ on C, then $\partial\Psi/\partial N = 0$ on C'. To remember (9), observe, heuristically, that $dN = |f'(z)|\, dn$ because $|f'(z)|$ is the local magnification (Exercise 1).

EXAMPLE 2. To illustrate (9), let us use the same problem as displayed in Fig. 1, but with the boundary conditions changed to Neumann boundary conditions: $\partial\psi/\partial n = 10e^{-x}$ on AB, $\partial\psi/\partial n = 30$ on EC, and $\partial\psi/\partial n = 25/(1+y)$ on FE.

Consider AB. With the help of (4), the $|f'(z)|$ in (9) is

$$\begin{aligned} |f'(z)| &= |2z| = 2\sqrt{x^2+y^2} = 2\sqrt{x^2+\frac{9}{x^2}} \\ &= 2\sqrt{\frac{u+\sqrt{u^2+36}}{2}+\frac{18}{u+\sqrt{u^2+36}}} \\ &= 2\sqrt[4]{u^2+36}. \end{aligned} \tag{10}$$

Further, the $\partial\psi/\partial n = 10e^{-x}$ needed in (9) is given by the right-hand side of (5), so

$$\frac{\partial\Psi}{\partial N} = \frac{1}{2\sqrt[4]{u^2+36}}\left[10\exp\left(-\sqrt{\frac{u+\sqrt{u^2+36}}{2}}\right)\right] \tag{11}$$

on $A'B'$. Next, consider EC. There, (2) gives $x = \sqrt{u}$, so

$$|f'(z)| = |2z| = 2x = 2\sqrt{u}$$

so (9) gives

$$\frac{\partial\Psi}{\partial N} = \frac{1}{2\sqrt{u}}(30) = \frac{15}{\sqrt{u}} \tag{12}$$

on $E'C'$. Finally, consider FE. We leave it for Exercise 2 to show that

$$\frac{\partial\Psi}{\partial N} = \frac{25}{2\left(\sqrt{-u}-u\right)} \tag{13}$$

on $F'E'$.

COMMENT. Simpler than (10) is the following approach:

$$|f'(z)| = |2z| = 2\,|z| = 2\sqrt{|w|} = 2\sqrt[4]{u^2+v^2}\Big|_{v=6} = 2\sqrt[4]{u^2+36}$$

on $A'B'$. ▮

Closure. The transformation of a Dirichlet boundary condition is simple: according to (1), $\Psi(u,v) = \psi(x,y)$. For instance, if $\psi = 47$ at a point z on the boundary C of the regon D, then $\Psi = 47$ at the image point w on the boundary C' of the region D'. Of course, one does need to determine the correspondence between points on C and points on C'. In Example 1, for example, to transform the boundary condition $\psi = 10e^{-x}$ on AB we need to solve for x (on AB) as a function of u (on $A'B'$).

In contrast, Neumann conditions do not carry over intact. For instance, if $\partial\psi/\partial n = 25$ at a point z on C, then $\partial\Psi/\partial N$ does not equal 25 at the image point w on C', it equals 25 times the scale factor $1/|f'(z)|$ which, in general, is not unity. That scale factor is due to the local magnification of dn: $dN = |f'(z)|\, dn$, as can be seen in the denominators in (9).

EXERCISES 22.5

1. In the sentence preceding Example 2, we stated that $|f'(z)|$ is the local amplification. That is, if $\Delta z = z - z_0$ is any vector springing from z_0, and $\Delta w = w - w_0$, where w_0 is the image of z_0 and w is the image of z, then the "magnification" $|\Delta w| / |\Delta z|$ is equal to $|f'(z_0)|$ in the limit as $\Delta z \to 0$ (z approaches z_0), if $w = f(z)$ is analytic at z_0 and $f'(z_0) \neq 0$ (i.e., if the map is conformal there). Why is that statement true?

2. Derive the expression (13) for $\partial\Psi/\partial N$ on $F'E'$.

3. In each case you are given the region D, Dirichlet and/or Neumann boundary conditions, and the mapping function $w(z)$. Determine the image D' and transformed conditions on its boundary. You need *not* solve for Ψ.

(a) Region D: $-\infty < x < \infty, 0 < y < 3\pi/2$. Boundary conditions: $\psi_y(x,0) = 5$, $\psi(x, 3\pi/2) = \sin x$. Map: $w(z) = e^z$.
(b) Region D: $-\infty < x < \infty, -\pi/2 < y < \pi/2$. Boundary conditions: $\psi_y(x, -\pi/2) = 20\cos x$, $\psi(x, \pi/2) = 3e^{-|x|}$. Map: $w(z) = e^z$.
(c) Region D: $0 < x < \infty, 0 < y < \pi$. Boundary conditions: $\psi_y(x,0) = e^{-x}$, $\psi_x(0,y) = 3y$, $\psi(x,\pi) = e^{-4x}$. Map: $w(z) = e^z$.
(d) Region D: $-\infty < x < 0, 0 < y < \pi/2$. Boundary conditions: $\psi(x,0) = 5e^x$, $\psi_x(0,y) = 2y$, $\psi_y(x,\pi/2) = 5$. Map: $w(z) = e^z$.
(e) Region D: $0 < x < 2, 0 < y < \pi$. Boundary conditions: $\partial\psi/\partial n = 10$ on all four edges. Map: $w(z) = e^z$.
(f) Region D: $0 < x < 1, 0 < y < \infty$. Boundary conditions: $\psi(0,y) = e^{-y}$, $\partial\psi/\partial n = 10x$ on $y = 0$, $\partial\psi/\partial n = 0$ on $x = 1$. Map: $w(z) = (z-1)/z$.
(g) Region D: $0 < x < \infty, -\infty < y < \infty$. Boundary conditions: $\psi(0,y) = 50$ on $-\infty < y \leq 0$, $\psi(0,y) = 50e^{-y}$ on $0 < y < \infty$. Map: $w(z) = (z-1)/(z+1)$.
(h) Region D: $0 < x < 1, 0 < y < \infty$. Boundary conditions: $\psi(0,y) = 100$, $\partial\psi/\partial n = -2$ on $y = 0$, $\psi(1,y) = 3e^{-y}$. Map: $w(z) = 1/z$.
(i) Region D: $0 < x < 1, -\infty < y < 0$. Boundary conditions: $\psi(0,y) = 25\cos y$, $\partial\psi/\partial n = 5$ on $y = 0$ and on $x = 1$. Map: $w(z) = 1/z$.
(j) Region D: $0 < x < \sqrt{1+y^2}, 0 < y < \infty$. Boundary conditions: $\psi(0,y) = \sin y$, $\partial\psi/\partial n = 1$ on $y = 0$, $\psi = 2e^{-y}$ on $x = \sqrt{1+y^2}$. Map: $w(z) = z^2$.
(k) Region D: $0 < x < \infty, 0 < y < \infty$. Boundary conditions: $\psi(0,y) = 5/(y^2+1)$, $\psi(x,0) = 2 - 5e^{-x}$. Map: $w(z) = z^2$.

22.6 Applications to Fluid Mechanics

Let $\mathbf{q}(x,y,z)$ be a fluid velocity field. If the flow is irrotational (i.e., $\nabla \times \mathbf{q} = \mathbf{0}$), then there exists a scalar velocity potential ϕ such that $\mathbf{q} = \nabla\phi$. Further, if the flow is incompressible (i.e., $\nabla \cdot \mathbf{q} = 0$), then $\nabla \cdot \nabla\phi = \nabla^2\phi = 0$, so ϕ satisfies the Laplace equation. Thus, an irrotational incompressible flow is called a *potential flow*. These points are discussed in Section 16.10. (There, we used $\mathbf{v}$ for the velocity field; here, it is more convenient to use $\mathbf{q}$.)

In the present section we consider only two-dimensional potential flows governed by

$$\nabla^2\phi = \phi_{xx} + \phi_{yy} = 0. \tag{1}$$

Throughout this section we assume that the boundary condition is $\mathbf{q}\cdot\hat{\mathbf{n}} = 0$ along any rigid boundaries, where $\hat{\mathbf{n}}$ is the usual outward normal vector (to the flow field). That is, the flow neither separates from the wall nor penetrates it. Thus, $\mathbf{q}\cdot\hat{\mathbf{n}} = \nabla\phi\cdot\hat{\mathbf{n}} = \partial\phi/\partial n = 0$, where the second equality amounts to the directional derivative formula $d\phi/ds = \nabla\phi\cdot\hat{\mathbf{s}}$ in any $\hat{\mathbf{s}}$ direction, from Section 16.4.

EXAMPLE 1. *Flow in a Corner.* Determine the potential flow in a corner, sketched in Fig. 1. The function $\zeta = z^2$ maps the first quadrant of the z plane, conformally, onto the upper half of the ζ plane, with the boundary condition $\partial\phi/\partial n = 0$ carrying over to $\partial\Phi/\partial N = 0$. Note that we use a $\zeta = \xi + i\eta$ plane, rather than the usual $w = u + iv$ plane because u, v designate the x, y fluid velocity components, here, and w will be used for the so-called complex velocity potential, defined below.

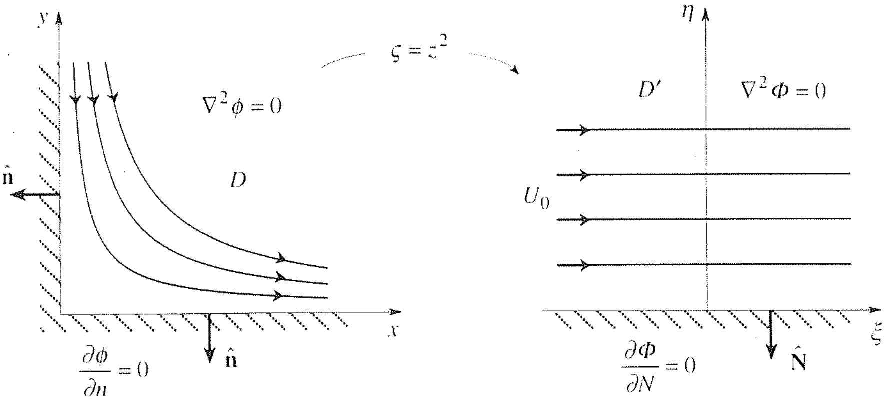


Figure 1. Corner flow.

A simple flow in the ζ plane, which will give the desired corner flow in the z plane, is the uniform stream shown in the figure, for any stream speed U_0. Then $\partial\Phi/\partial\xi = U_0$ and $\partial\Phi/\partial\eta = 0$, so $\Phi(\xi,\eta) = U_0\xi$. Since

$$\zeta = \xi + i\eta = (x+iy)^2 = x^2 - y^2 + i2xy, \tag{2}$$

we have $\xi = x^2 - y^2$ and $\eta = 2xy$ so

$$\phi(x,y) = U_0\xi(x,y) = U_0(x^2 - y^2). \tag{3}$$

Hence, the desired velocity field is given by

$$\mathbf{q} = \nabla\phi = 2U_0(x\hat{\mathbf{i}} - y\hat{\mathbf{j}}). \tag{4}$$

Recall, from our study of the gradient in Section 16.4, that at each point in the field the vector $\nabla\phi$ is normal to the $\phi =$ constant curve through that point. It is also true that if $\psi(x,y)$ is the conjugate harmonic function corresponding to $\phi(x,y)$ so that $w(z) =$

$\phi(x,y)+i\psi(x,y)$ is an analytic function of z, then the ψ = constant curves are orthogonal to the ϕ = constant curves (as is discussed in Exercise 16 in Section 21.5). Thus, the fluid particles move along ψ = constant curves. These curves are called *streamlines*, and $\psi(x,y)$ is called the *stream function*. (For additional discussion of the stream function, see the exercises in Section 16.10.)

Corresponding to $\phi(x,y) = U_0(x^2 - y^2)$, the conjugate harmonic function $\psi(x,y)$ is readily found (Exercise 1) to be

$$\psi(x,y) = 2U_0xy \tag{5}$$

so the streamlines $\psi(x,y) = 2U_0xy$ = constant are the hyperbolas xy = constant as shown in Fig. 1. ▮

Observe that we solved the potential problem for $\phi(x,y)$. From ϕ, we obtained the velocity field as $\mathbf{q} = \nabla\phi$, and we obtained the streamline pattern by finding the conjugate harmonic function $\psi(x,y)$ corresponding to $\phi(x,y)$. However, in the fluid mechanics literature it is more common to work with the **complex velocity potential**,

$$w(z) = \phi(x,y) + i\psi(x.y), \tag{6}$$

right from the start, rather than ϕ. Then

$$\frac{dw}{dz} = \phi_x + i\psi_x = \phi_x - i\phi_y = u - iv, \tag{7}$$

where the first equality holds because $w(z)$ is analytic (see Theorem 22.5.1), and the second one holds because ϕ and ψ satisfy the Cauchy–Riemann conditions. Thus, we recover the x, y velocity components from $w(z)$ as $u = \operatorname{Re} w'(z)$ and $v = -\operatorname{Im} w'(z)$, respectively.

Let us denote

$$w(z(\zeta)) = W(\zeta) = \Phi(\xi,\eta) + i\Psi(\xi,\eta). \tag{8}$$

Then, by the same reasoning as used in (7),

$$\frac{dW}{d\zeta} = U - iV, \tag{9}$$

where $U(\xi,\eta)$ and $V(\xi,\eta)$ are the ξ, η velocity components, respectively.

To illustrate the use of the complex potential w, let us rework Example 1 using w. In the ζ plane we have

$$\frac{dW}{d\zeta} = U(\xi,\eta) - iV(\xi,\eta) = U_0 - i0 \tag{10}$$

so $W = U_0\zeta$ (plus an arbitrary constant, which can be set equal to zero without loss). Returning to the z plane via the mapping function $\zeta = z^2$, we have

$$w = U_0z^2 = U_0(x^2 - y^2) + i2U_0xy \tag{11}$$

and

$$\frac{dw}{dz} = 2U_0 z = 2U_0 x + i2U_0 y, \tag{12}$$

so the quantities of interest are $\phi(x,y) = U_0(x^2 - y^2)$, $\psi(x,y) = 2U_0 xy$, $u(x,y) = 2U_0 x$, and $v(x,y) = -2U_0 y$, as found in Example 1. Notice, in particular, that $u = v = 0$ at the corner ($x = y = 0$), which is therefore a *stagnation point.*

We consider just one more example, to further illustrate the use of the complex potential and to introduce the important Joukowski transformation.

EXAMPLE 2. *Flow Over Semicircular Bump.* We seek the plane potential flow over a semicircular bump of radius a, where the flow tends to a uniform stream $\mathbf{q} = U_0\hat{\mathbf{i}}$ as $r \to \infty$ along any $\theta =$ constant ray (Fig. 2). The boundary condition is that $\partial\phi/\partial n = 0$ all along the solid boundary $ABCEF$.

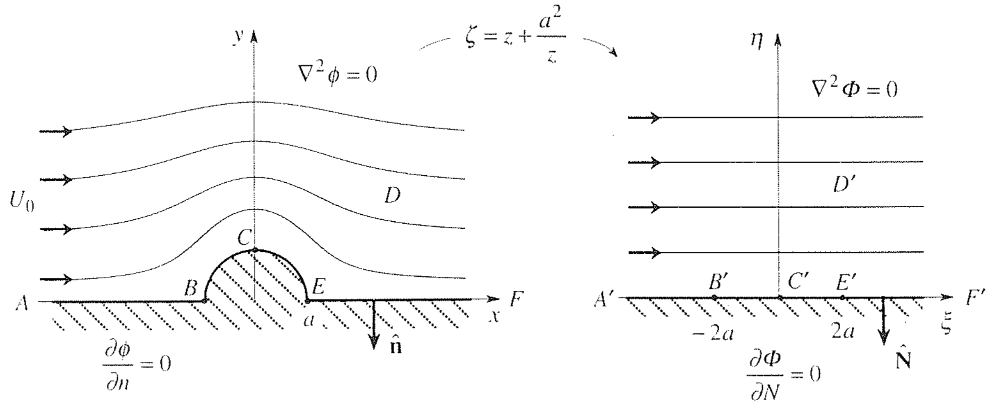


Figure 2. Flow over semicircular bump.

From Appendix F we find that the **Joukowski transformation**

$$\boxed{\zeta = z + \frac{a^2}{z}} \tag{13}$$

maps D, conformally, onto the upper half of the ζ plane (Fig. 2). Besides preserving the homogeneous Neumann boundary condition (because it is a conformal map), the Joukowski transformation also preserves the boundary condition at infinity,

$$\frac{dw}{dz} = u(x,y) - iv(x,y) \sim U_0 \qquad \text{as } z \to \infty, \tag{14}$$

because (13) gives $\zeta \sim z$ as $z \to \infty$. That is, it follows from

$$\frac{dw}{dz} = \frac{dW}{d\zeta}\frac{d\zeta}{dz} \sim \frac{dW}{d\zeta}, \tag{15}$$

and (14), that

$$\frac{dW}{d\zeta} \sim U_0 \qquad \text{as } \zeta \to \infty. \tag{16}$$

In fact, the flow in the ζ plane is simply a free stream

$$W = U_0\zeta, \tag{17}$$

not just at infinity but everywhere. Finally, putting (13) into (17) gives the desired complex potential

$$w = U_0\left(z + \frac{a^2}{z}\right). \tag{18}$$

Since the boundary of D is made up of constant r and constant θ curves it is convenient to put $z = re^{i\theta}$ (rather than $z = x + iy$) into (18) so

$$w = U_0\left(r + \frac{a^2}{r}\right)\cos\theta + iU_0\left(r - \frac{a^2}{r}\right)\sin\theta. \tag{19}$$

Thus,

$$\phi(r,\theta) = U_0\left(r + \frac{a^2}{r}\right)\cos\theta, \tag{20a}$$

$$\psi(r,\theta) = U_0\left(r - \frac{a^2}{r}\right)\sin\theta, \tag{20b}$$

and

$$\frac{dw}{dz} = u - iv = U_0\left(1 - \frac{a^2}{z^2}\right) = U_0\left(1 - \frac{a^2}{r^2}\cos 2\theta\right) + iU_0\frac{a^2}{r^2}\sin 2\theta \tag{21}$$

so the velocity components are

$$u(r.\theta) = U_0\left(1 - \frac{a^2}{r^2}\cos 2\theta\right), \tag{22a}$$

$$v(r,\theta) = -U_0\frac{a^2}{r^2}\sin 2\theta. \tag{22b}$$

COMMENT 1. Observe that (22) does satisfy the boundary condition at infinity: $u \sim U_0$ and $v \to 0$ as $r \to \infty$. And from (20b) we see that the solid boundary $ABCEF$ (Fig. 2) is indeed a streamline, namely $\psi = 0$: on EF and AB the $\sin\theta$ factor is zero, and on BCE the $r - a^2/r$ is zero.

COMMENT 2. The closed form solution (20a) to the problem shown in Fig. 2 could also have been obtained by separation of variables.

COMMENT 3. We pointed out that the Joukowski transformation (13) preserves the flow at infinity, from the z plane to the ζ plane, because (13) gives $\zeta \sim z$ as $z \to \infty$. It is interesting that we can think of (13) as the simplest (nontrivial) case of the more general transformation

$$\zeta = z + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots$$

having that property. ▮

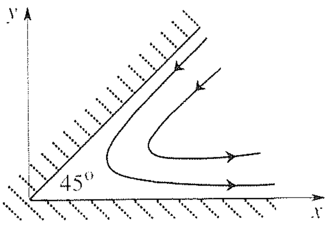
Closure. The key idea in this section is the use of the complex velocity potential $w(z) = \phi(x,y) + i\psi(x,y)$, rather than just $\phi(x,y)$. Once we find w, the x and

y velocity components, u and v, are found from $dw/dz = \phi_x + i\psi_x = u - iv$, and the streamlines are found from $\operatorname{Im} w = \psi(x,y) =$ constant. We also meet the Joukowski transformation (13), which has the property of preserving the conditions at infinity because it simply gives $\zeta \sim z$ as $z \to \infty$. We will have more to say about that transformation in the exercises.

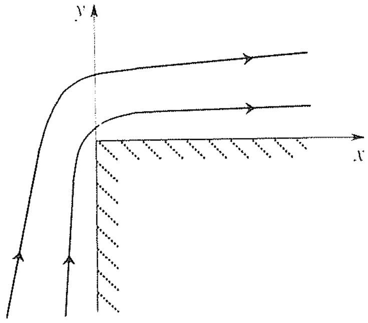
EXERCISES 22.6

1. Given $\phi(x,y) = U_0(x^2 - y^2)$, derive the conjugate harmonic function given by (5).

2. (a) Instead of the 90° corner shown in Fig. 1, consider the 45° corner shown here. Solve for the corner flow sketched in the figure. That is, solve for $w(z)$, $\phi(x,y)$, $\psi(x,y)$, $u(x,y)$, and $v(x,y)$. Show whether or not the corner is a stagnation point, as it was for the 90° corner flow in Example 1.



(b) The same as part (a) but for the "outside corner" shown.



(c) The same as part (a) but for any corner angle α (radians). [In (a), for instance, $\alpha = \pi/4$.]
(d) In view of the results in part (a), would you say that the corner flow found in Example 1 is unique? Explain. HINT: A flow is possible for which the 45° ray $y = x$ is a streamline. Another flow is possible for which 30° and 60° rays are streamlines. And so on.

3. Use computer plotting software, such as the *Maple* implicitplot command, to obtain the z plane flow pattern shown in Fig. 2, using four or five representative streamlines, somewhat as we have.

4. In each case find $w(z)$, $\phi(x,y)$, $\psi(x,y)$, $u(x,y)$, and $v(x,y)$.

(a) Solve for the downward and rightward potential flow between the positive x and y axes and the curve $xy = 6$ in the first quadrant.
(b) Solve for the downward potential flow in the upper half plane, between $xy = -1$, $xy = 4$, and the x axis.

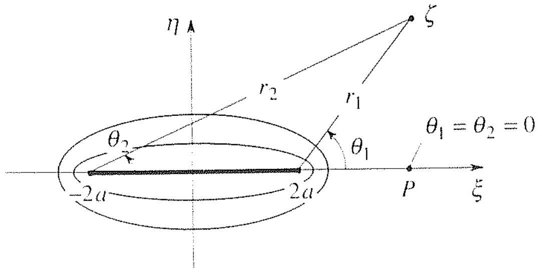
5. (*Joukowski transformation*) Show that the Joukowski transformation

$$\xi + i\eta = z + \frac{a^2}{z} \qquad (a > 0) \tag{5.1}$$

maps the family of circles $x^2 + y^2 = c^2$ onto the confocal ellipses

$$\frac{\xi^2}{(c + a^2/c)^2} + \frac{\eta^2}{(c - a^2/c)^2} = 1 \tag{5.2}$$

with foci at $\zeta = \pm 2a$ as shown in the figure. In particular,

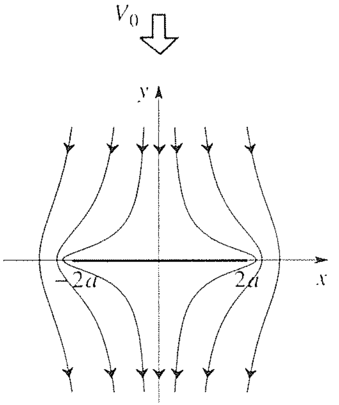


show that the circle $|z| = a$ maps onto the line segment $\eta = 0$, $|\xi| \le 2a$. Show, further, that the region $|z| > a$ maps onto the entire ζ plane minus that line segment, and that the region $|z| < a$ does too, so that the inverse transformation is double-valued. In fact, show that

$$z = \frac{\zeta + \sqrt{\zeta^2 - 4a^2}}{2}, \tag{5.3}$$

which is indeed double-valued because of the square root. Finally, show that with $\sqrt{\zeta^2 - 4a^2}$ defined by the branch cut shown in the figure (and extending from $-2a$ to $+2a$), the image of the slit ζ plane is $|z| > a$, whereas if we change the branch cut by defining $\theta_1 = 0$ and $\theta_2 = 2\pi$ at P, then the image is $|z| < a$. NOTE: For discussion of the finite branch cut, see the optional Section 21.4.7.

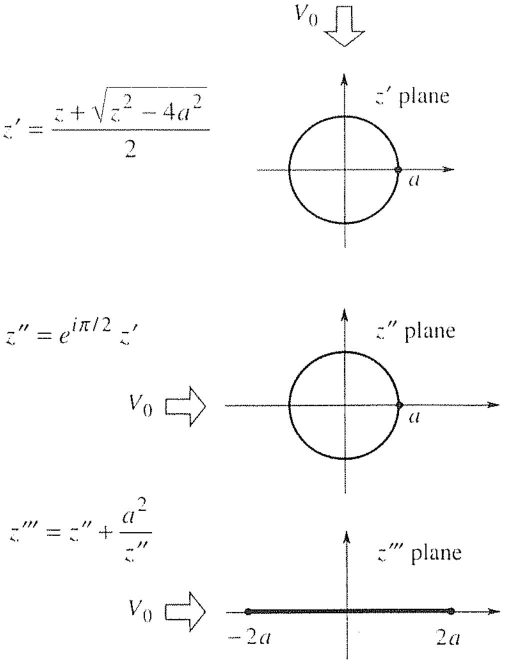
6. (*Flow perpendicular to flat plate*) (a) Solve for the potential flow around a thin flat plate that is perpendicular to a uniform stream V_0; that is, $dw/dz = u - iv \sim iV_0$ as $z \to \infty$. (Actually, that flow is not unique, but it is if we ask it to be symmetric about the y axis, which symmetry we assume.) Specifically,



solve for $w(z)$. HINT: Use the mapping sequence suggested in the figure showing the z', z'', and z''' planes.
(b) Show that on the top of the plate $u = V_0 x/\sqrt{4a^2 - x^2}$ and $v = 0$, and on the bottom of the plate $u = -V_0 x/\sqrt{4a^2 - x^2}$ and $v = 0$.
(c) Show that $\psi(x, y)$ can be expressed in terms of x and y as

$$\psi(x, y) = \frac{V_0}{2}\left[\sqrt{(r_1 + r_1\cos\theta_1)(r_2 + r_2\cos\theta_2)} - \sqrt{(r_1 - r_1\cos\theta_1)(r_2 - r_2\cos\theta_2)}\right], \tag{6.1}$$

where $r_1 = \sqrt{(x - 2a)^2 + y^2}$, $r_2 = \sqrt{(x + 2a)^2 + y^2}$, $r_1\cos\theta_1 = x - 2a$, and $r_2\cos\theta_2 = x + 2a$. HINT: $\cos(A + B) = \cos A\cos B - \sin A\sin B$, $\cos\frac{A}{2} = \sqrt{(1 + \cos A)/2}$, $\sin\frac{A}{2} = \sqrt{(1 - \cos A)/2}$.



(d) Use computer plotting software, such as the *Maple* implicitplot command, to obtain the flow pattern in the z plane. (Take $a = 1$, say.)

Chapter 22 Review

A conformal map $w(z) = f(z) = u(x, y) + iv(x, y)$ is a change of variables from x, y to u, v, that simplifies the given two-dimensional potential problem by simplifying the domain, while preserving the governing Laplace partial differential equation. For $w = f(z)$ to be conformal, we need $f(z)$ to be analytic and $f'(z) \neq 0$ everywhere on the domain. Besides simplifying the domain while preserving the Laplace equation, such maps are conformal in the sense that they preserve angles, both in magnitude and sense, which property is often useful.

A virtue of conformal mapping is that it gives solutions in closed form. A disadvantage of the method, however, is that there does not exist a systematic procedure that inevitably leads us to a suitable mapping function $f(z)$. Rather, one needs to rely on some familiarity with various important maps such as the bilinear transformation, Joukowski transformation, some of the elementary functions, and so on. Fortunately, there exist extensive tables of conformal maps.

Thus, our program consists mainly of studying a number of different maps – the bilinear map in Section 22.3, and additional maps in Section 22.4. We noted that the bilinear map, which has the useful property of sending circles into circles, is the only map that is one-to-one so that, in general, one needs to be involved in selecting suitable branches of multi-valued functions.

Considering only the simplest boundary conditions through Section 22.4, constant boundary conditions of Dirichlet type, we show how to handle nonconstant Dirichlet conditions, and Neumann conditions as well, in Section 22.5.

Finally, in Section 22.6 we look specifically at applications of conformal mapping to problems in fluid mechanics, that case being just slightly different in that it is convenient, and traditional, to work with the complex potential $\phi(x, y) + i\psi(x, y)$, the z derivative of which gives $u - iv$, where u and v are the x, y velocity components, respectively.

Chapter 23

The Complex Integral Calculus

23.1 Introduction

In Chapters 21 and 22 we got as far as the differential calculus of functions of a complex variable, with special emphasis on the concept of analyticity. After applying the analytic function theory to the solution of two-dimensional potential problems by conformal mapping, in Chapter 22, we now pick up where we left off in Chapter 21 and turn to the complex integral calculus. We begin by defining the complex integral $\int_C f(z)\,dz$, and then proceed to derive three major integral theorems: the *Cauchy theorem*, the *fundamental theorem of the complex integral calculus*, and the *Cauchy integral formula*.

From the Cauchy integral formula we will derive *Taylor series* for functions of a complex variable, and a generalization of the latter known as *Laurent series*, in Chapter 24. Laurent series will enable us to categorize singularities into different types and to thereby understand them better. It also leads us, in the final section of Chapter 24, to the important *residue theorem* of the complex integral calculus.

It would be natural to expect the evaluation of complex integrals to be more difficult than the evaluation of real integrals. Thus, it is surprising to discover that it is often very simple, thanks to the extremely powerful integral theorems mentioned above.

23.2 Complex Integration

23.2.1. Definition and properties. Having studied the Riemann integral $\int_a^b f(x)\,dx$ of a real valued function f on a real x axis, in the calculus, we generalized that concept to line integrals in two or three dimensions, in Chapter 16. Here, we extend the concept once more and introduce the **complex integral**

$$I = \int_C f(z)\,dz \tag{1}$$

of a given function f along a given oriented curve C in the complex z plane; f may be analytic or not, and C may be either a closed curve or an arc, but in either case we shall assume that C is piecewise smooth and simple.* The curve C is called the *path of integration*, or *contour*. To define the integral (1), we stay as close as possible to the definition of line integrals developed in Chapter 16.

Suppose that C has initial and final points $z = A$ and $z = B$, respectively. If C is closed, then $B = A$. Divide C into n arcs by specifying points $z_0 = A$, $z_1, z_2, \dots, z_{n-1}, z_n = B$ along C. Let the division be chosen arbitrarily, provided that the points z_j are spaced and numbered so that the arc length from A to z_j is less than the arc length from A to z_k if $j < k$ (Fig. 1). Denote the n arcs as $C_1, C_2, \dots, C_n$, where the endpoints of C_j are z_{j-1} and z_j. On each C_j choose some point Q_j that is anywhere between the endpoints of C_j or at one of the endpoints, and form the sum

$$J_n = \sum_{j=1}^{n} f(Q_j)\Delta z_j, \tag{2}$$

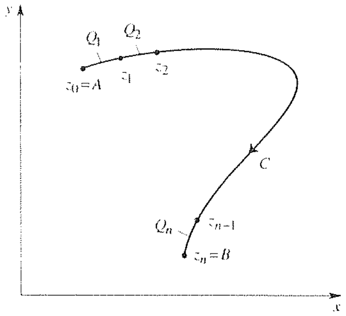


Figure 1. Partition of C.

where $\Delta z_j = z_j - z_{j-1}$. The choice of the z_j's and Q_j's defines a *partition* of C, and we call the largest $|\Delta z_j|$ the *norm* of the partition. We introduce not just one partition but a sequence of them such that the norm of the nth partition tends to zero as $n \to \infty$. If the corresponding sequence of sums $J_1, J_2, \dots$ converges to a limit, we call that limit the integral $\int_C f(z)\,dz$. We then say that the integral exists – that is, that it is *convergent*.

From this definition it can be shown that complex integration is *linear* – that is,

$$\int_C [\alpha f(z) + \beta g(z)]\,dz = \alpha \int_C f(z)\,dz + \beta \int_C g(z)\,dz \tag{3}$$

for any scalars α and β and for any functions f and g for which the three integrals exist, and that if we break C into two parts, C_1 and C_2 (Fig. 2), then

$$\int_C f(z)\,dz = \int_{C_1} f(z)\,dz + \int_{C_2} f(z)\,dz. \tag{4}$$

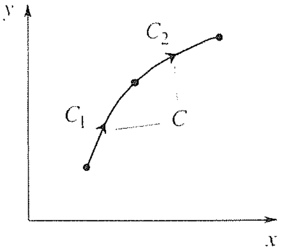


Figure 2. Breaking up the path.

Finally, if we reverse the orientation of C and call the reverse path "$-C$," we have

$$\int_{-C} f(z)\,dz = -\int_C f(z)\,dz. \tag{5}$$

EXAMPLE 1. *Using the Limit Definition to Evaluate Integrals.* To illustrate the limit definition of the complex integral, consider the simple case where $f(z) = z$. Since we can choose each Q_j point anywhere along its C_j arc, let us choose each at the beginning of its arc: Q_1 at z_0, Q_2 at z_1, and so on. Then (2) becomes

$$J_n^b = z_0(z_1 - z_0) + z_1(z_2 - z_1) + \cdots + z_{n-2}(z_{n-1} - z_{n-2}) + z_{n-1}(z_n - z_{n-1}), \tag{6}$$

*These terms are defined in Section 15.2.

where the superscript b simply denotes that the sum corresponds to choosing each z_j at the beginning of its C_j arc. We can just as well choose each Q_j at the end of its C_j interval, which choice gives

$$J_n^e = z_1(z_1 - z_0) + z_2(z_2 - z_1) + \cdots + z_{n-1}(z_{n-1} - z_{n-2}) + z_n(z_n - z_{n-1}). \quad (7)$$

Adding (6) and (7), most terms cancel, leaving

$$\begin{aligned} J_n^b + J_n^e &= z_n^2 - z_0^2 \\ &= B^2 - A^2. \end{aligned} \quad (8)$$

Finally, refining the partition over and over so that its norm tends to zero, the left-hand side of (8) tends to $2I$ so $I = (B^2 - A^2)/2$, or

$$I = \int_C z\,dz = \left.\frac{z^2}{2}\right|_A^B, \quad (9)$$

where A and B are the initial and final points of C, respectively. ▮

Why all the fuss? Doesn't $\int_C z\,dz = (z^2/2)|_A^B$ follow immediately from the fact that $d(z^2/2)/dz = z$? Yes it does, but we have not yet developed the fundamental theorem of the complex integral calculus, on which result that claim would be based; all we have, thus far, is the limit definition.

Of course, the limit definition is too unwieldy to be useful in evaluating more complicated integrals so we proceed to develop other lines of approach. In this regard it will be illuminating to distinguish two basically different strategies.

One strategy is to reduce $\int_C f(z)\,dz$ to one or more *real* integrals. To do so, one merely needs to re-express

$$\begin{aligned} \int_C f(z)\,dz &= \int_C (u + iv)(dx + i\,dy) \\ &= \int_C (u\,dx - v\,dy) + i\int_C (v\,dx + u\,dy). \end{aligned} \quad (10)$$

Thus, we can evaluate $\int_C f(z)\,dz$ by evaluating the two real line integrals

$$\int_C [u(x,y)\,dx - v(x,y)\,dy] \qquad \text{and} \qquad \int_C [v(x,y)\,dx + u(x,y)\,dy]$$

by methods developed earlier in Chapter 16. Let us illustrate with two examples.

EXAMPLE 2. Evaluate

$$I = \int_C z^2 dz, \quad (11)$$

where C is the parabolic arc shown in Fig. 3. Parametrizing C according to $y = \tau$, $x = 4 - \tau^2$, as τ goes from $+2$ to -2, we have

$$\begin{aligned}
I &= \int_C [(x^2 - y^2)dx - 2xy\,dy] + i\int_C [2xy\,dx + (x^2 - y^2)dy] \\
&= \int_2^{-2} \left[(x^2 - y^2)\frac{dx}{d\tau} - 2xy\frac{dy}{d\tau}\right] d\tau + i\int_2^{-2} \left[2xy\frac{dx}{d\tau} + (x^2 - y^2)\frac{dy}{d\tau}\right] d\tau \\
&= \int_2^{-2} \left\{\left[(4 - \tau^2)^2 - \tau^2\right](-2\tau) - 2(4 - \tau^2)\tau\right\} d\tau \\
&\qquad + i\int_2^{-2} \left\{2(4 - \tau^2)\tau(-2\tau) + \left[(4 - \tau^2)^2 - \tau^2\right]\right\} d\tau \\
&= 0 + i\,\frac{16}{3} = \frac{16}{3}\,i. \qquad (12)
\end{aligned}$$

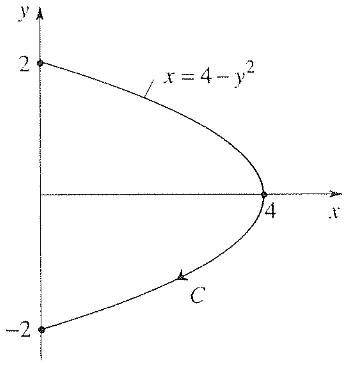


Figure 3. The path C in (11).

COMMENT. It is natural to wonder whether we could have simplified the evaluation of I by deforming C into a simpler shape such as a straight line from the initial point $2i$ to the final point $-2i$. This same important question arose in Chapter 16 when we studied line integrals, and we found that path deformation (between fixed endpoints) is permissible if the given vector field is sufficiently well behaved. In view of the connection between $\int_C f(z)\,dz$ and real line integrals, displayed in (10), we expect that the same is true for the complex integral $\int_C f(z)\,dz$: the path C can be deformed if $f(z)$ is sufficently well behaved. We defer discussion of this matter to Section 23.3. ▮

EXAMPLE 3. Evaluate

$$I = \oint_C (z - a)^n dz, \qquad (13)$$

where a is a given complex number, n is any integer (positive, negative, or zero), and C is a circle of radius R, centered at $z = a$ and oriented counterclockwise (Fig. 4). In this case C is *closed*. Generally, we will use the $\oint_C$ notation when the contour C is closed.

It is convenient to use the polar angle ϕ to parametrize C, by setting

$$z - a = Re^{i\phi}. \qquad (14)$$

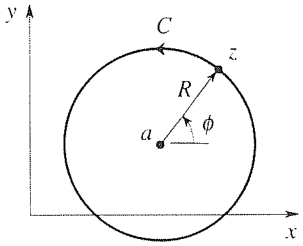


Figure 4. The closed contour C in (13).

That is, the parametric equations of C are, from the real and imaginary parts of (14), $x = \operatorname{Re} a + R\cos\phi$ and $y = \operatorname{Im} a + R\sin\phi$, as ϕ increases from 0 to 2π. Then

$$\begin{aligned}
I &= \int_0^{2\pi} (Re^{i\phi})^n (Rie^{i\phi}\,d\phi) = iR^{n+1}\int_0^{2\pi} e^{i(n+1)\phi}\,d\phi \\
&= \frac{R^{n+1}}{n+1}\,e^{i(n+1)\phi}\Big|_0^{2\pi} = 0,
\end{aligned}$$

provided that $n \neq -1$ (in which case the latter yields the indeterminate form $0/0$). Treating $n = -1$ separately,

$$I = iR^0 \int_0^{2\pi} e^{i0\phi}\,d\phi = i\int_0^{2\pi} d\phi = 2\pi i.$$

Thus

$$\oint_C (z-a)^n\, dz = \begin{cases} 2\pi i, & n=-1 \\ 0, & n \neq -1 \end{cases} \tag{15}$$

so $I = 0$ for all n's except $n = -1$, in which case $I = 2\pi i$. ∎

The other possible strategy is to keep the complex integral intact and to develop theorems that will enable us to work with $\int_C f(z)\, dz$ directly. It will be found that this approach is by far the more fruitful and will be pursued exclusively for the remainder of this chapter and the next. Looking ahead, we shall obtain Cauchy's theorem, the fundamental theorem of the integral calculus, Cauchy's integral formula, and the residue theorem, in that order. These integral theorems will prove powerful indeed, and will also yield additional important results regarding function theory in general.

In that development we will often need to be able to "bound" complex integrals, that is, to obtain bounds on their absolute magnitude (modulus). Thus, in the remainder of this section we consider a simple upper bound that will suffice for most of our needs.

23.2.2. Bounds. First, observe from (2) that

$$|J_n| = \left| \sum_{j=1}^{n} f(Q_j)\Delta z_j \right| \leq \sum_{j=1}^{n} |f(Q_j)\Delta z_j| = \sum_{j=1}^{n} |f(Q_j)|\, |\Delta z_j|, \tag{16}$$

where the inequality and the final equality follow from Exercise 5(f) and 5(d), respectively, in Section 21.2. If there exists a real constant M such that $|f(z)| \leq M$ everywhere on C, then $|f(Q_j)| \leq M$ for each j, and (16) gives

$$|J_n| \leq M \sum_{j=1}^{n} |\Delta z_j|. \tag{17}$$

As $n \to \infty$ and the norm of the sequence of partitions tends to zero, the sum on the right-hand side tends to L, the length of the (presumably rectifiable) curve C. Thus, the final result is the bound

$$\boxed{\left| \int_C f(z)\, dz \right| \leq ML,} \tag{18}$$

where $|f(z)| \leq M$ on C and L is the length of C. Subsequently, we shall refer to this result as the the **ML bound**.

EXAMPLE 4. Use (18) to bound the integral

$$I = \int_C \frac{e^z}{z^2}\, dz, \tag{19}$$

where C is the straight line shown in Fig. 5. To obtain M, in (18), write

$$|f(z)| = \left|\frac{e^z}{z^2}\right| = \frac{|e^{x+iy}|}{|z^2|} = \frac{|e^x|\,|e^{iy}|}{|z|^2} = \frac{e^x}{x^2+y^2}. \tag{20}$$

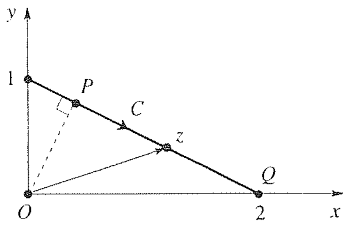


Figure 5. The contour C in (19).

On C, e^x is a maximum at Q and x^2+y^2 is a minimum at P so we expect the maximum of $|f(z)|$ to lie somewhere between P and Q. To find that point we put $y = 1 - \frac{1}{2}x$ into $e^x/(x^2+y^2)$, differentiate with respect to x, and set the derivative equal to zero. Those steps give $x = 4/5$. With $x = 4/5$ and $y = 3/5$, (20) gives $M = e^{4/5}/(\frac{16}{25}+\frac{9}{25}) = e^{4/5}$. And since $L = \sqrt{5}$, (18) gives

$$|I| \le e^{4/5}\sqrt{5} = 4.976. \tag{21}$$

COMMENT. In this example we actually found the maximum value of $|f(z)|$ on C, and used that as M. In most applications, however, finding the maximum value of $|f(z)|$ on C is so difficult that we are willing to accept a cruder value of M – that is, a bound on $|f(z)|$ that is greater than the maximum value of $|f(z)|$. In the present example, for instance, the greatest value of the numerator is e^2 (at Q), and the smallest value of the denominator is the distance OP squared, which, by similar triangles, is $(2/\sqrt{5})^2 = 4/5$, from which it follows that

$$|I| \le \left(\frac{e^2}{4/5}\right)\left(\sqrt{5}\right) = 20.65. \tag{22}$$

The trade-off is that (22) was more easily obtained than (21), but is somewhat cruder; we say that the inequality (22) is not as "sharp" as (21).[*] ▮

One more example:

EXAMPLE 5. Use (18) to bound the integral

$$I = \int_C \frac{\sin z}{z(z^2+9)}\,dz, \tag{23}$$

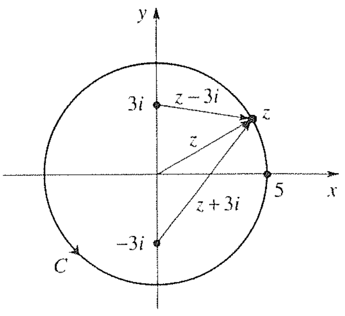


Figure 6. The contour C in (23).

where C is the circular contour shown in Fig. 6. In the spirit of the comment in Example 4, let it suffice to seek an upper bound on the numerator, $\sin z$, and a lower bound on the denominator, $z(z^2+9)$. On C we have

$$\begin{aligned}
|\sin z| &= |\sin(x+iy)| \\
&= |\sin x \cosh y + i \cos x \sinh y| \\
&= \sqrt{\sin^2 x \cosh^2 y + \cos^2 x \sinh^2 y} \\
&\le \sqrt{\cosh^2 y + \sinh^2 y} \\
&\le \sqrt{\cosh^2 5 + \sinh^2 5} = \sqrt{\cosh 10},
\end{aligned} \tag{24}$$

[*] The less sharp, the less informative. For instance, to say that the author owns less than a million neckties is less informative than saying that he owns less than seven, which is still less informative than saying that he owns exactly five.

$$|z| = 5, \tag{25}$$

$$\begin{aligned} |z^2 + 9| &= |(z - 3i)(z + 3i)| = |z - 3i|\,|z + 3i| \\ &\geq (2)(2) = 4, \end{aligned} \tag{26}$$

where the second inequality in (24) follows from the fact that $\cosh y$ and $\sinh y$ are monotonically increasing functions of y, and the last equality follows from the identity $\cosh(A + B) = \cosh A \cosh B + \sinh A \sinh B$ (Exercise 8 of Section 21.3). Further, the minimum values of $|z - 3i|$ and $|z + 3i|$ occur at $z = +5i$ and $-5i$, respectively, and equal 2. With $L = 2\pi(5) = 10\pi$, it follows from (24)–(26) and (18) that

$$|I| \leq \left(\frac{\sqrt{\cosh 10}}{(5)(4)}\right)(10\pi) = \frac{\pi\sqrt{\cosh 10}}{2} \tag{27}$$

is a bound on I. ▮

Closure. Beginning with the limit definition of the complex integral $I = \int_C f(z)\,dz$, we note that (as in the case of real integrals) that definition is not useful for the evaluation of such integrals, except in the simplest cases. As an alternative, we show that we can obtain the form $I = \int_C (u\,dx - v\,dy) + i\int_C (v\,dx + u\,dy)$, where the two line integrals can be evaluated by whatever methods were developed for line integrals in Chapter 16. But we promised more powerful approaches – the fundamental theorem of the complex integral calculus, Cauchy's theorem, the Cauchy integral formula, and the residue theorem, which are developed over the remainder of this chapter and the next.

We also show how to obtain a simple upper bound on I, the "ML bound" $|I| \leq ML$, which is needed in the subsequent sections.

EXERCISES 23.2

1. Evaluate the following by expressing them in terms of real line integrals and then evaluating those integrals.

(a) $\int_C |z|^2\,dz$, where C is a straight line from $z = 0$ to $z = 1 + i$
(b) $\int_C \bar{z}\,dz$, where C is the same as in part (a)
(c) $\int_C \bar{z}\,dz$, where C is a clockwise semicircle from $z = 2$ to $z = -2$, centered at $z = 0$
(d) $\int_C dz/z$, where C consists of three straight-line segments: from $z = 1$ to $z = 1 - i$, from $z = 1 - i$ to $z = -1 - i$, and then from $z = -1 - i$ to $z = -1$
(e) $\int_C e^z\,dz$, where C consists of two straight-line segments: from $z = i$ to $z = 1 + i$, and then from $z = 1 + i$ to $z = 1 - 2i$
(f) $\int_C (\text{Re}\,z)dz$, where C is a clockwise quarter circle from $z = 3i$ to $z = 3$, centered at $z = 0$
(g) $\int_C (\text{Im}\,z)dz$, where C is a straight line from $z = i$ to $z = 2 + 2i$

2. Consider $I = \int_C \bar{z}\,dz$, where the initial and final points of C are $z = 0$ and $z = 1 + i$, respectively. Show that the integral is *path dependent* by choosing two different paths and obtaining different values for I.

3. Use (18) to obtain an upper bound on $|I|$ for each given integral I. NOTE: Be aware that the answer is not unique, because we seek an inequality rather than an equality, just as both of the bounds (21) and (22) were correct in Example 4. Thus, it will be important for you to show your steps and reasoning.

(a) $\int_C z^5\,dz$, where C is a straight line from $z = i$ to $z = 2+2i$
(b) $\int_C e^z\,dz$, where C is a straight line from $z = -2$ to $z = 1 + 3i$
(c) $\int_C e^{-z}\,dz$, where C is the same as in part (b)

(d) $\int_C dz/z$, where C is a circle of radius 4 centered at $z = 2 + i$
(e) $\int_C dz/(z^2 + 1)$, where C is a circle of radius 3 centered at $z = 0$
(f) $\int_C (z^2 + 1)dz/(z^2 - 1)$, where C is a circle of radius 1 centered at $z = 1$
(g) $\int_C dz/[z(z + i)]$, where C is a quarter circle from $z = i$ to $z = 1$, centered at $z = 0$
(h) $\int_C e^z \, dz/z$, where C is a straight line from $z = i$ to $z = 2$
(i) $\int_C \cos z \, dz/z$, where C is the same as in part (h)

4. Obtain, by any means, the maximum value of $|\sin z/[z(z^2 + 9)]|$, in Example 5, on the contour C shown in Fig. 6, correct to two significant figures.

5. (a) If C is a straight line from $z = 2$ to $z = 2 + (\pi/2)i$, show that

$$\left| \int_C \frac{dz}{e^z + 1} \right| \leq \frac{\pi}{2} \frac{1}{\sqrt{e^4 + 1}}.$$

(b) If C is a straight line from $z = 2i$ to $z = 3$, show that

$$\left| \int_C \frac{\cos z}{z} \, dz \right| \leq \frac{13}{6} \cosh 2.$$

6. In our subsequent work we will sometimes need an upper bound on the absolute magnitude of a given integral, and (18) will generally suffice. A corresponding *lower* bound was not developed in this section because it will not be needed. Nevertheless, suppose that $|f(z)| \geq m$ on C. It is tempting to conjecture, by analogy with (18), that

$$\left| \int_C f(z) \, dz \right| \geq mL \tag{6.1}$$

provides a lower bound on the magnitude of the integral. Prove that (6.1) is *not* correct.

7. (*A sharper bound than the ML bound*) (a) Derive the upper bound

$$\boxed{\left| \int_C f(z) \, dz \right| \leq \int_C |f(z)| \, |dz| \, ,} \tag{7.1}$$

and show that

$$\left| \int_C f(z) \, dz \right| \leq \int_C |f(z)| \, |dz| \leq ML, \tag{7.2}$$

from which it follows that (7.1) is, in general, sharper than the ML bound [i.e., closer to an equality than (18)]. NOTE: The price that we pay for this improvement is that (7.1) is generally more difficult to apply than (18) because it requires the evaluation of the (real) integral $\int_C |f(z)| \, |dz|$. For our purposes, in later sections, (18) will almost always suffice; only in one or two cases will the more refined bound (7.1) be needed.
(b) To better understand these two bounds, apply each of them to the integral $I = \int_0^2 (x^3 - x)dx$, that is, $\int_C (z^3 - z)dz$ where C is a straight line from $z = 0$ to $z = 2$. Compare the two results with the exact value of I. Further, show the graphical significance of each of the three members of (7.2), by suitable labeled sketches.

23.3 Cauchy's Theorem

Recall that the contour of integration may be open or closed. Cauchy's theorem, which we now derive, involves closed paths. Specifically, let C be a piecewise smooth simple closed curve, and express

$$\oint_C f(z) \, dz = \oint_C (u \, dx - v \, dy) + i \oint_C (u \, dy + v \, dx). \tag{1}$$

Suppose that $f(z)$ is analytic and that $f'(z)$ is continuous in a simply connected domain D containing the path C. If we write the first integral on the right-hand side of (1) as

$$\oint_C (u \, dx - v \, dy) = \oint_C \mathbf{w} \cdot d\mathbf{R}, \tag{2}$$

where $\mathbf{w} = u(x,y)\hat{\mathbf{i}} - v(x,y)\hat{\mathbf{j}}$ is a vector field and $d\mathbf{R} = dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}$ (where the z in dz is the real variable z, not $z = x + iy$), then

$$\nabla \times \mathbf{w} = 0\hat{\mathbf{i}} - 0\hat{\mathbf{j}} + \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right)\hat{\mathbf{k}} \tag{3}$$

vanishes everywhere in D by virtue of the Cauchy–Riemann equation $\partial u/\partial y = -\partial v/\partial x$ (which holds because f has been assumed analytic in D). Then it follows from Theorem 16.10.1 that

$$\oint_C (u\,dx - v\,dy) = 0. \tag{4}$$

Similarly, if we write the second integral on the right-hand side of (1) as

$$\oint_C (u\,dy + v\,dx) = \oint_C \mathbf{w} \cdot d\mathbf{R}, \tag{5}$$

where this time $\mathbf{w} = v(x,y)\hat{\mathbf{i}} + u(x,y)\hat{\mathbf{j}}$, then

$$\nabla \times \mathbf{w} = 0\hat{\mathbf{i}} - 0\hat{\mathbf{j}} + \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)\hat{\mathbf{k}} \tag{6}$$

vanishes everywhere in D by virtue of the other Cauchy – Riemann equation, $\partial u/\partial x = \partial v/\partial y$. Then it follows that

$$\oint_C (u\,dy + v\,dx) = 0. \tag{7}$$

Thus, we have shown that *if $f(z)$ is analytic and $f'(z)$ is continuous in a simply connected domain D, then*

$$\oint_C f(z)\,dz = 0 \tag{8}$$

for every piecewise smooth simple closed curve C in D.

This result was published by Cauchy in 1825. Seventy-five years later it was shown by *Edouard Goursat* (1858 – 1936) that the stated result holds even if one does not assume the continuity of $f'(z)$ (i.e., of u_x, u_y, v_x, v_y).* Indeed, we will show (in Section 23.5) that if f is analytic in D (i.e., once differentiable), then f possesses derivatives of *all* orders in D. In particular, since f' is differentiable in D, it *follows* that f' must be continuous there.

Thus, deleting the assumption that $f'(z)$ is continuous, we have the so-called "strong version" of Cauchy's theorem, sometimes called the Cauchy–Goursat theorem in recognition of Goursat's contribution:

*For proof, see, for example, R. V. Churchill, *Complex Variables and Applications*, 2nd ed. (New York: McGraw – Hill, 1960), Chap.5, or Konrad Knopp, *Theory of Functions*, Part I (New York: Dover, 1945), Chap. 4.

THEOREM 23.3.1 *Cauchy's Theorem*
If $f(z)$ is analytic in a simply connected domain D, then

$$\boxed{\oint_C f(z)\,dz = 0} \tag{9}$$

for every piecewise smooth simple closed curve C in D.

EXAMPLE 1. Consider

$$I = \oint_C \frac{dz}{z^2 - 5z + 6} = \oint_C \frac{dz}{(z-2)(z-3)}, \tag{10}$$

where C is the unit circle, traversed counterclockwise (Fig. 1). The integrand $f(z) = 1/[(z-2)(z-3)]$ is analytic everywhere except at $z = 2$ and $z = 3$. Thus, it is analytic in the simply connected domain D (Fig. 1) containing C so $I = 0$ by Cauchy's theorem. ▮

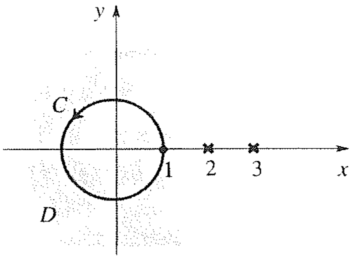


Figure 1. C and D for Example 1.

We have already mentioned the connection between the complex integral $\int_C f(z)\,dz$ of an analytic function f and the real line integral $\int_C \mathbf{w} \cdot d\mathbf{R}$ of an irrotational field $\mathbf{w}$. Referring once more to Theorem 16.10.1, notice the equivalence between the vanishing of $\int_C \mathbf{w} \cdot d\mathbf{R}$ for every closed loop C and the path independence of integrals $\int_C \mathbf{w} \cdot d\mathbf{R}$ between fixed endpoints. Thus, we might well anticipate such equivalence for complex integrals as well. In fact, we do have the following corollary to Theorem 23.3.2:

THEOREM 23.3.2 *Path Independence*
If $f(z)$ is analytic in a simply connected domain D, then $\int_C f(z)\,dz$ is independent of path in D. That is, given any initial point P in D and any final point Q in D, the value of $\int_C f(z)\,dz$ is the same for every piecewise smooth path C, lying entirely within D, from P to Q.

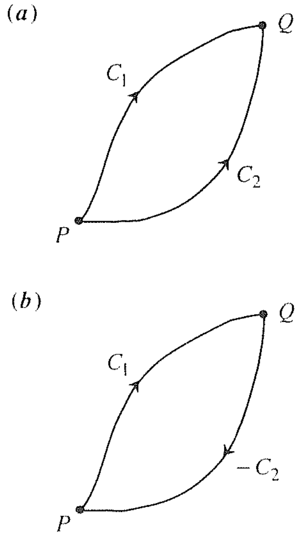


Figure 2. You take the high road, I'll take the low.

Partial Proof: Let C_1 and C_2 be any two piecewise smooth paths from P to Q. Suppose that both C_1 and C_2 are simple curves and that they intersect each other only at the endpoints P and Q (Fig. 2a). (The additional argument needed for the case where intersections do occur, possibly an infinite number of them, is omitted here.) Then $C_1 + (-C_2)$ is a piecewise smooth simple closed curve in D (Fig. 2b) so, according to Cauchy's theorem,

$$\oint_{C_1 + (-C_2)} f(z)\,dz = 0. \tag{11}$$

But

$$\oint_{C_1+(-C_2)} f(z)\,dz = \int_{C_1} f(z)\,dz + \int_{-C_2} f(z)\,dz$$
$$= \int_{C_1} f(z)\,dz - \int_{C_2} f(z)\,dz, \tag{12}$$

and it follows from (11) and (12) that

$$\int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz,$$

as claimed. This step completes the proof, except for the omission noted above. ■

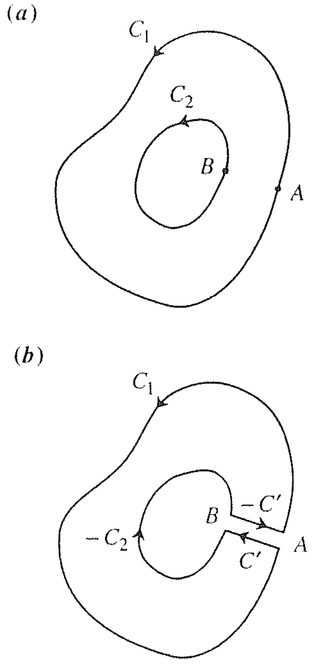


Figure 3. Path deformation for closed path.

It is useful to think of path independence in terms of a process of *path deformation*. That is, we can imagine deforming C_1, continuously, into C_2 (Fig. 2a), keeping the endpoints fixed. If f is analytic on C_1 and C_2 and we cross no singular points in the process (thus f is analytic on and between C_1 and C_2), then $\int_{C_1} f(z)\,dz$ is equal to $\int_{C_2} f(z)\,dz$. The point, of course, is that we may be able to simplify the integral by such deformation in the sense that $\int_{C_2} f(z)\,dz$ may be more easily evaluated than $\int_{C_1} f(z)\,dz$.

In fact, path deformation can be applied to closed paths as well. For example, suppose that f is analytic on and between the closed paths C_1 and C_2 shown in Fig. 3a. Let us "slit" the region by introducing a piecewise smooth curve connecting a point A on C_1 with a point B on C_2 (with this curve intersecting C_1 only at A and C_2 only at B). Denote this curve as C' when it is oriented from A to B and as $-C'$ when it is oriented from B to A as shown in Fig. 3b. (Actually, C' and $-C'$ are coincident, but we have separated them in the figure for clarity.) Since $C = C_1 + C' + (-C_2) + (-C')$ is a piecewise smooth simple closed curve and f is analytic inside and on C, it follows from Cauchy's theorem that $\oint_C f(z)\,dz = 0$ (see Exercise 3). Thus,

$$\oint_C f(z)\,dz = \oint_{C_1+C'+(-C_2)+(-C')} f(z)\,dz$$
$$= \oint_{C_1} f(z)\,dz + \int_{C'} f(z)\,dz + \oint_{-C_2} f(z)\,dz + \int_{-C'} f(z)\,dz$$
$$= \oint_{C_1} f(z)\,dz + \int_{C'} f(z)\,dz - \oint_{C_2} f(z)\,dz - \int_{C'} f(z)\,dz$$
$$= 0, \tag{13}$$

so

$$\oint_{C_1} f(z)\,dz = \oint_{C_2} f(z)\,dz. \tag{14}$$

That is, deforming C_1 into C_2, (14) will hold if f is analytic between and on C_1 and C_2.

EXAMPLE 2. *An Important Little Integral.* Consider

$$I = \oint_C (z-a)^n \, dz, \tag{15}$$

where C is any piecewise smooth simple closed curve, oriented counterclockwise, such as the curve C shown in Fig. 4; n is any integer (positive, negative, or zero), and a is any given complex number lying inside C. According to the discussion preceding this example, we can deform the contour C to a *circular* contour C', lying wholly within C, as depicted in Fig. 4 since $(z-a)^n$ is analytic on and between C and C'. The resulting integral, $\oint_{C'}(z-a)^n \, dz$, is evaluated in Example 3 in Section 23.2. Recalling the result of that evaluation, we have

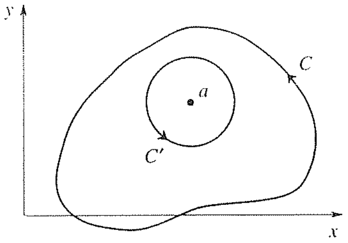


Figure 4. Deformation of C to a circle C'.

$$\boxed{\oint_C (z-a)^n \, dz = \begin{cases} 2\pi i, & n=-1 \\ 0, & n \neq -1. \end{cases}} \tag{16}$$

That is, the result obtained in Example 3 of Section 23.2 holds even if C is not a circle centered at a. Although rather specific, formula (16) will be used often enough so it is worth remembering. ▮

EXAMPLE 3. Evaluate

$$I = \oint_C \frac{dz}{z^2(z-2)(z-4)}, \tag{17}$$

where C is given in Fig. 5. Expanding the integrand in partial fractions yields

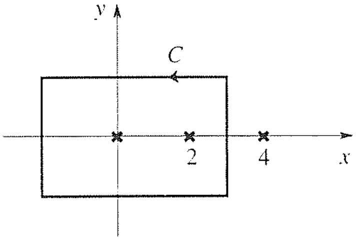


Figure 5. The contour C in (17).

$$I = \frac{3}{32}\oint_C \frac{dz}{z} + \frac{1}{8}\oint_C \frac{dz}{z^2} - \frac{1}{8}\oint_C \frac{dz}{z-2} + \frac{1}{32}\oint_C \frac{dz}{z-4}. \tag{18}$$

Thus

$$I = \frac{3}{32}(2\pi i) + \frac{1}{8}(0) - \frac{1}{8}(2\pi i) + \frac{1}{32}(0) = -\frac{\pi i}{16}, \tag{19}$$

because the last integral in (18) is zero by Cauchy's theorem, and the first three are covered by the "important little integral" (16). ▮

Closure. Roughly put, Cauchy's theorem tells us that $\oint_C f(z)\,dz = 0$ if $f(z)$ is analytic within a region containing C, and it follows from that result that if C_1 and C_2 are open curves with the same initial point and the same final point, then $\int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz$ if $f(z)$ is analytic within a region containing C_1 and C_2. That is, we can deform C_1 into C_2 without changing the value of the integral. These results, Cauchy's theorem and its path independence corollary, are fundamental to the further development of the complex integral calculus in this chapter and the next. The connection between them and the irrotational field theorem (Theorem 16.10.1) should be understood.

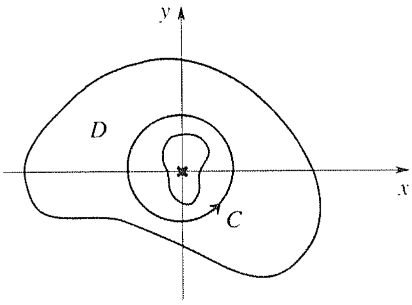
EXERCISES 23.3

1. According to Example 2,

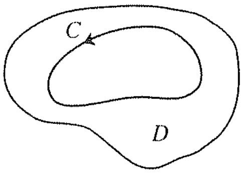
$$\oint_C \frac{dz}{z^2} = 0,$$

where C is a counterclockwise circle of radius R, centered at the origin. Yet $f(z) = 1/z^2$ is not analytic within C; it is singular at $z = 0$. Explain why this result does not violate Cauchy's theorem.

2. Consider $I = \oint_C dz/z$, where C is the counterclockwise unit circle, and the assertion that $I = 0$, from Cauchy's theorem, because $f(z) = 1/z$ is analytic in the domain D containing C. (See the accompanying figure.) Yet we show in Example 2 that $I = 2\pi i$. Explain the apparent contradiction.



3. Prove that Cauchy's theorem can be stated a bit more simply, as follows: *If C is a piecewise smooth simple closed curve and $f(z)$ is analytic inside and on C, then $\oint_C f(z)\,dz = 0$.* HINT: The idea is to show that if f is analytic inside and on C, then there does exist a domain D (see the figure) containing C, such that f is analytic within D. The desired result follows immediately from Cauchy's theorem. If needed, you may draw on the fact that C is necessarily rectifiable (i.e., of finite length) because it is piecewise smooth. Also, recall from the calculus that if $F(x)$ is continuous on a closed interval $a \le x \le b$, then it has an absolute maximum and an absolute minimum on that interval.



4. Let C_1, C_2, C_3 be the following simple closed curves:

C_1: $|z| = 1$, counterclockwise
C_2: $|z| = 1$, clockwise
C_3: the square with vertices at $1 - i$, $1 + i$, $-1 + i$, $-1 - i$, counterclockwise.

Evaluate each of the following integrals using Cauchy's theorem if applicable, or any other method studied in Sections 23.2 or 23.3.

(a) $\oint_{C_1} \operatorname{Re} z\, dz$
(b) $\oint_{C_1} \operatorname{Im} z\, dz$
(c) $\oint_{C_3} \operatorname{Im} z\, dz$
(d) $\oint_{C_3} \dfrac{dz}{z^2 - 3}$
(e) $\oint_{C_1} \dfrac{dz}{z^4}$
(f) $\oint_{C_1} \dfrac{dz}{z(z-2)}$
(g) $\oint_{C_2} \dfrac{dz}{z(z+5)}$
(h) $\oint_{C_1} e^{\sin z}\, dz$
(i) $\oint_{C_2} \sin(\cos z)\, dz$
(j) $\oint_{C_3} \dfrac{dz}{|z|}$
(k) $\oint_{C_1} \bar{z}\, dz$
(l) $\oint_{C_3} \bar{z}\, dz$

5. Can we use path deformation to obtain

$$\oint_{C_3} \bar{z}\, dz = \oint_{C_1} \bar{z}\, dz, \tag{5.1}$$

where C_1 and C_3 are defined in Exercise 4? Explain.

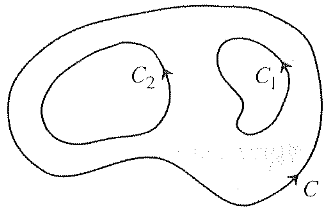
6. Evaluate $\int_C z^{20}\, dz$, where C is the path

(a) $y = x - x^3$, from $x = 0$ to $x = 1$
(b) $y = x - x^3$, from $x = 0$ to $x = -1$
(c) $x = y - y^5$, from $y = 0$ to $y = 1$

7. Evaluate $\int_C \bar{z}\, dz$, where C is a straight line from $z = 0$ to $z = 1 + i$, where C is the parabola $y = x^2$ from $z = 0$ to $z = 1 + i$, and where C is the rectilinear path from $z = 0$ to $z = 1$ to $z = 1 + i$. Are the answers the same? Is there any violation of Theorem 23.3.2? Explain.

8. (*Path deformation in multiply-connected domain*) Show that if $f(z)$ is analytic in the shaded region between and on the contours C, C_1, C_2 (see the accompanying figure), then

$$\oint_C f(z)\,dz = \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz. \tag{8.1}$$



You may use the result stated in Exercise 3. NOTE: More generally,

$$\oint_C f(z)\,dz = \oint_{C_1} f(z)\,dz + \cdots + \oint_{C_n} f(z)\,dz$$

if $C_1, \dots, C_n$ are nonintersecting counterclockwise closed contours within C, and f is analytic between and on $C, C_1, \dots, C_n$.

9. Evaluate the following integrals, where in each case C is the circle $|z| = 3$, counterclockwise.

(a) $\oint_C \dfrac{dz}{z(z-1)}$ (b) $\oint_C \dfrac{dz}{z(z-5)}$

(c) $\oint_C \dfrac{z\,dz}{z^2+1}$ (d) $\oint_C \dfrac{z\,dz}{z^2-3z+2}$

(e) $\oint_C \dfrac{dz}{z^3(z^2-1)}$ (f) $\oint_C \dfrac{z\,dz}{z^3-1}$

23.4 Fundamental Theorem of the Complex Integral Calculus

Before obtaining the fundamental theorem of the complex integral calculus, let us, for the sake of comparison and motivation, recall the *fundamental theorem of the real integral calculus*: let $f(x)$ be continuous on the interval $x_0 \le x \le x_1$. Then the function

$$G(x) = \int_{x_0}^{x} f(\xi)\,d\xi \qquad (x_0 \le x \le x_1) \tag{1}$$

is differentiable on $x_0 \le x \le x_1$ and is a *primitive* or *indefinite integral* of f; that is,

$$G'(x) = f(x) \tag{2}$$

on that closed interval. The primitive of f is unique to within an arbitrary additive constant. Finally, if $F(x)$ is any particular primitive of f, then

$$\int_{x_0}^{x} f(\xi)\,d\xi = F(x) - F(x_0). \tag{3}$$

The foregoing result is found, in the elementary calculus, to be especially valuable for evaluating integrals for which a primitive can be found by inspection. For example, to evaluate $\int_2^5 x^2\,dx$ one notices that $x^3/3$ is a primitive of x^2 [because $d(x^3/3)/dx = x^2$] so

$$\int_2^5 x^2\,dx = \frac{x^3}{3}\bigg|_2^5 = \frac{125-8}{3} = \frac{117}{3}.$$

Of course, the primitive $(x^3/3) + 4$, say, would have produced the same result.

Turning to the complex case, let $f(z)$ be analytic in a simply-connected domain D. If z_0 is a fixed point in D, then $\int_{z_0}^{z} f(\zeta)\, d\zeta$ is (Theorem 23.3.2) path independent and hence defines a single-valued function of z,* which we denote as $G(z)$:

$$G(z) = \int_{z_0}^{z} f(\zeta)\, d\zeta. \tag{4}$$

(Before continuing, notice that in the complex case we have asked f to be analytic, for the integral to define a single-valued function, whereas in the real case it sufficed to ask f to be continuous.)

Next, we show that $G'(z) = f(z)$ in D, analogous to equation (2) for the real case. Consider the difference quotient

$$\begin{aligned}\frac{G(z+\Delta z) - G(z)}{\Delta z} &= \frac{1}{\Delta z}\left[\int_{z_0}^{z+\Delta z} f(\zeta)\, d\zeta - \int_{z_0}^{z} f(\zeta)\, d\zeta\right]\\ &= \frac{1}{\Delta z}\int_{z}^{z+\Delta z} f(\zeta)\, d\zeta\\ &= \frac{1}{\Delta z}\int_{z}^{z+\Delta z} f(z)\, d\zeta + \frac{1}{\Delta z}\int_{z}^{z+\Delta z} [f(\zeta) - f(z)]\, d\zeta\\ &= f(z) + \frac{1}{\Delta z}\int_{z}^{z+\Delta z} [f(\zeta) - f(z)]\, d\zeta, \end{aligned} \tag{5}$$

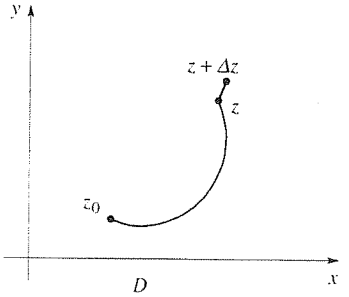


Figure 1. z_0, z, $z+\Delta z$, and the path.

where z_0, z, and $z+\Delta z$ are displayed in Fig. 1. By the above-noted path independence, there is no loss in taking the path from z to $z + \Delta z$ to be straight (although the path from z_0 to z may need to be curved in order to remain within D). Using the ML bound on the last integral in (5), we have

$$\left|\frac{1}{\Delta z}\int_{z}^{z+\Delta z} [f(\zeta) - f(z)]\, d\zeta\right| \le \frac{1}{|\Delta z|} M\, |\Delta z| = M, \tag{6}$$

where $M = \max|f(\zeta) - f(z)|$ on the line from z to $z+\Delta z$. Since f is continuous (because it is analytic), $M \to 0$ as $\Delta z \to 0$. Thus, letting $\Delta z \to 0$ in (5) gives

$$G'(z) = f(z), \tag{7}$$

as claimed.

Any function $F(z)$ satisfying $F'(z) = f(z)$ is called an **indefinite integral** or **primitive** of f. It is easy to show (Exercise 1) that any two primitives corresponding to a given f differ at most by an arbitrary additive constant. Thus, if $F(z)$ is any particular primitive of $f(z)$, then

$$G(z) = \int_{z_0}^{z} f(\zeta)\, d\zeta = F(z) + C, \tag{8}$$

*This is a key point. Recall that a function is to be single-valued. If f is not analytic in D then there is no guarantee (because Theorem 23.3.2 calls for analyticity) that it will be single-valued. See Example 2.

where C is a constant. To evaluate C, set $z = z_0$. Then

$$0 = F(z_0) + C \tag{9}$$

gives $C = -F(z_0)$ so

$$\int_{z_0}^{z} f(\zeta)\, d\zeta = F(z) - F(z_0), \tag{10}$$

analogous to equation (3) for the real case.

Pulling these results together, we have:

THEOREM 23.4.1 *Fundamental Theorem of the Complex Integral Calculus*
Let $f(z)$ be analytic in a simply-connected domain D, and let z_0 be any fixed point in D. Then

(i) $G(z) = \int_{z_0}^{z} f(\zeta)\, d\zeta$ is analytic in D and $G'(z) = f(z)$.

(ii) If $F(z)$ is any primitive of $f(z)$ [i.e., $F'(z) = f(z)$], then

$$\boxed{\int_{z_0}^{z} f(\zeta)\, d\zeta = F(z) - F(z_0).} \tag{11}$$

As in the analogous case of real integrals, (11) is especially useful for evaluating integrals for which a primitive can be found by inspection.

EXAMPLE 1. To evaluate $\int_{2i}^{3} \sin z\, dz$, observe that $F(z) = -\cos z$ (plus any constant) is a primitive of $\sin z$. Then

$$\int_{2i}^{3} \sin z\, dz = -\cos z \Big|_{z=2i}^{z=3} = -\cos 3 + \cos 2i = \cosh 2 - \cos 3. \tag{12}$$

COMMENT. Notice that the notation $\int_{2i}^{3}$ suffices. That is, we do not need to specify the path to be followed, from $2i$ to 3, because $\sin z$ is analytic in the whole plane, and hence line integrals of $\sin z$ are path independent – they depend only on the endpoints. ▮

EXAMPLE 2. Evaluate

$$I = \int_{1+i}^{-i} \frac{dz}{z}. \tag{13}$$

With the foregoing comment in mind, we observe, first, that the problem is not clearly posed, because the integral is *not* single-valued. Specifically, the integrand $1/z$ is analytic for all $z \neq 0$, and the whole plane with $z = 0$ deleted is a multiply connected region (due to the hole at $z = 0$). Thus our path independence theorem (Theorem 23.3.2) does not guarantee that I will have the same values for paths such as C_1 and C_2 in Fig. 2. In fact,

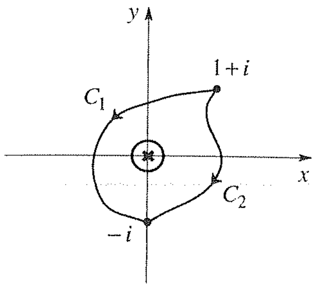


Figure 2. Alternative paths.

$$\int_{C_1} \frac{dz}{z} - \int_{C_2} \frac{dz}{z} = \int_{C_1+(-C_2)} \frac{dz}{z} = 2\pi i, \tag{14}$$

according to the "important little integral" in Section 23.3 so

$$\int_{C_1} \frac{dz}{z} \neq \int_{C_2} \frac{dz}{z}. \tag{15}$$

Sure enough, this multi-valuedness of the integral shows up if we write

$$I = \int_{1+i}^{-i} \frac{dz}{z} = \log z \Big|_{1+i}^{-i}, \tag{16}$$

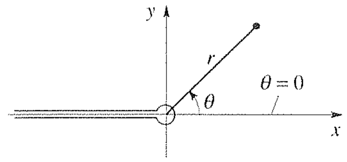


Figure 3. Branch cut for log z.

because $\log z$ is multi-valued! To render I single-valued we need to render the $\log z$ single-valued, and we do that, as usual, by a branch cut. If, for instance, we choose to adopt the branch cut shown in Fig. 3, then (16) gives

$$I = \log\left(re^{i\theta}\right)\Big|_{1+i}^{-i} = (\ln r + i\theta)\Big|_{r=\sqrt{2},\, \theta=\pi/4}^{r=1,\, \theta=-\pi/2} = -\frac{\ln 2}{2} - \frac{3\pi}{4}\, i \tag{17}$$

as the unique value of I. ▮

Closure. The purpose of this brief section is to establish the fundamental theorem of the complex integral calculus, Theorem 23.4.1. Although analogous to the familiar real variable result, we do note that whereas the real variable theorem merely asks the integrand $f(x)$ to be continuous, the complex variable theorem asks $f(z)$ to be analytic. As a simple illustration of Theorem 23.4.1, we can say that

$$\int_0^z \zeta^2\, d\zeta = \frac{\zeta^3}{3}\Big|_0^z = \frac{z^3}{3}$$

for all z, because $d(z^3/3)/dz = z^2$ is analytic for all z.

EXERCISES 23.4

1. Prove the assertion, stated below equation (7), that "any two primitives corresponding to a given f differ at most by an arbitrary additive constant." HINT: Let $F_1(z)$ and $F_2(z)$ be primitives of $f(z)$ so that $F_1'(z) = f(z)$ and $F_2'(z) = f(z)$. Subtract the latter two equations.

2. Give three different primitives for each of the following functions.

(a) z (b) z^5 (c) $e^{2z} - z$ (d) $\cos(z-2)$

3. Use the fundamental theorem to evaluate each of the following.

(a) $\int_0^i z\, dz$ (b) $\int_{-i}^i z^4\, dz$

(c) $\int_4^{-1-2i} (e^{-z} - 3z^2)\, dz$ (d) $\int_i^0 \cos 3z\, dz$

(e) $\int_{1-i}^{1+i} ze^z\, dz$ (f) $\int_4^{-i} z \sin z\, dz$

(g) $\int_0^{3i} ze^{z^2}\, dz$ (h) $\int_0^{1+2i} \sin^2 z\, dz$

(i) $\int_{-2i}^3 z \cos 2z\, dz$ (j) $\int_i^{2+i} \cosh 3z\, dz$

(k) $\int_0^i \cos^3 z\, dz$ (l) $\int_i^1 z^2 e^{2z^3}\, dz$

4. Determine all possible values of

$$I = \int_{1-i}^{1+i} \frac{dz}{z(z-1)}.$$

5. At some point in the calculus one is likely to encounter the calculation

$$I = \int_{-\infty}^{\infty} \frac{dx}{x^2+1} = \lim_{\substack{A \to \infty \\ B \to \infty}} \tan^{-1} x \Big|_{x=-A}^{x=B} = \pi. \tag{5.1}$$

Surely the integral is uniquely determined, so it is interesting that the multi-valued $\tan^{-1}(\,)$ enters the picture. However, no matter which continuous branch of $\tan^{-1}(\,)$ is chosen [e.g., $-\pi/2 < \tan^{-1}(\,) < \pi/2$, $\pi/2 < \tan^{-1}(\,) < 3\pi/2$, etc.] the unique result, π, is indeed obtained. For study purposes, it may be useful to reexamine this evaluation in the light of complex variable theory. Then

$$I = \int_C \frac{dz}{z^2+1} = \lim_{\substack{A \to \infty \\ B \to \infty}} \tan^{-1} z \Big|_{z=-A}^{z=B}, \tag{5.2}$$

where C is a straight line path from $-A$ to B along the real axis. To evaluate the right-hand member of (5.2), show that

$$\tan^{-1} z = \frac{1}{2i} \log \frac{i-z}{i+z}, \tag{5.3}$$

so that

$$I = \int_C \frac{dz}{z^2+1} = \lim_{\substack{A \to \infty \\ B \to \infty}} \frac{1}{2i} \log \frac{i-z}{i+z} \Bigg|_{z=-A}^{z=B} \tag{5.4}$$

[or, equivalently, derive (5.4) by applying the method of partial fractions to the integral]. Finally, evaluate the right-hand member of (5.4) by introducing any suitable branch cuts for the $\log(i-z)$ and $\log(i+z)$ functions, and show that you do obtain π, as in (5.1).

23.5 Cauchy Integral Formula

Recall that Cauchy's theorem is based upon analyticity of the integrand. Our purpose in the present section is to begin our consideration of the more typical case, where the integrand is singular at one or more points within the contour.

Let $f(z)$ be analytic in a simply connected domain D, let C be any piecewise smooth simple closed curve in D, oriented counterclockwise, and consider the integral

$$I = \oint_C \frac{f(z)}{z-a}\, dz, \tag{1}$$

where a is any fixed point within C (Fig. 1). That is, rather than a vague statement such as "Consider $\oint_C f(z)\, dz$ where $f(z)$ is singular somewhere within C," we consider the integrand to have a specific kind of singularity [namely, a $1/(z-a)$ behavior, which "blows up" as $z \to a$], and we make its presence explicit by writing the integrand as $f(z)/(z-a)$. Other types of singularity will be considered later.

To evaluate I, we begin by deforming C to a circular contour C' of radius ρ, sufficiently small so that C' lies entirely inside* C, as shown in Fig. 1. The

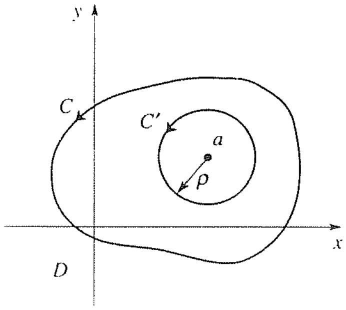


Figure 1. C in (1). D chosen as rectangular for simplicity.

*We often speak of the region *inside* or *outside* a given closed curve. Although there is little chance of misunderstanding in this regard, we wish to point out that to render the notion of inside and outside precise is not at all simple, and there exists a sophisticated result known as the **Jordan curve theorem** which clarifies the matter in a rigorous way. Discussion of this theorem lies well outside our present scope and can be found in R. N. Pederson, "The Jordan Curve Theorem for Piecewise Smooth Curves," *American Mathematical Monthly*, Vol. 76, 1969, pp. 605–610, or in G. N. Watson, *Complex Integration and Cauchy's Theorem*, Cambridge Tracts No. 15, 1914, Chap. 1.

deformation is justified since the integrand $f(z)/(z-a)$ is analytic between and on C and C'.

Understand that $I = \oint_{C'} f(z)\,dz/(z-a)$ is independent of ρ, provided of course that C' stays within D. Thus, we can let $\rho \to 0$, which is convenient because then $f(z) \sim f(a)$ on C' and the integral simplifies to $I = \oint_{C'} f(a)\,dz/(z-a) = f(a)\oint_{C'} dz/(z-a) = f(a)(2\pi i)$ since $\oint_{C'} dz/(z-a) = 2\pi i$ according to our "important little integral." To prove this result rigorously, re-express (1) as

$$\begin{aligned} I &= \oint_C \frac{f(z)}{z-a}\,dz = \oint_{C'} \frac{f(z)}{z-a}\,dz \\ &= \oint_{C'} \frac{f(a)}{z-a}\,dz + \oint_{C'} \frac{f(z)-f(a)}{z-a}\,dz \\ &= f(a)(2\pi i) + \oint_{C'} \frac{f(z)-f(a)}{z-a}\,dz. \end{aligned} \tag{2}$$

The third equality is simply an identity, rigged so that the first integral on the right is the anticipated final result and the second integral is a "deviation term." By letting $\rho \to 0$ in (2), we expect to be able to show that the deviation term is zero. Specifically, the ML bound gives

$$\left|\oint_{C'} \frac{f(z)-f(a)}{z-a}\,dz\right| \le \frac{M}{\rho}2\pi\rho = 2\pi M, \tag{3}$$

where $M = \max|f(z)-f(a)|$ on C'. Since $f(z)$ is continuous (because it is analytic), $M \to 0$ as $\rho \to 0$ so, letting $\rho \to 0$ in (2), we obtain $I = 2\pi i f(a)$, which result is known as the **Cauchy integral formula**.*

THEOREM 23.5.1 *Cauchy Integral Formula*
Let $f(z)$ be analytic in a simply-connected domain D, let C be a piecewise smooth simple closed curve in D oriented counterclockwise, and let a be any point within C (Fig. 1). Then

$$\boxed{\oint_C \frac{f(z)}{z-a}\,dz = 2\pi i f(a).} \tag{4}$$

Let us illustrate the use of (4) in evaluating integrals.

EXAMPLE 1. Evaluate

$$I = \oint_C \frac{e^z}{(z-2)(z+4)}\,dz, \tag{5}$$

where C is a counterclockwise circle of radius 3, centered at the origin (Fig. 2).

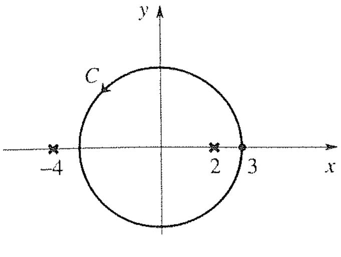


Figure 2. The contour C in (5).

*It would be misleading to say that the deviation term tends to zero as $\rho \to 0$, misleading because it is not a function of ρ. Rather, put it this way: we *find out* that the deviation is zero by letting $\rho \to 0$.

The first step is to examine the integrand to see where, if anywhere, it is singular. Evidently it has two singular points, one at $z = 2$ and one at $z = -4$. Of these, only $z = 2$ falls inside C so, comparing the left-hand sides of (4) and (5), we can identify

$$f(z) = e^z/(z+4) \qquad \text{and} \qquad a = 2.$$

Thus, (4) gives

$$I = 2\pi i \left(\frac{e^z}{z+4} \right) \bigg|_{z=2} = \frac{\pi e^2}{3} i. \tag{6}$$

Of course, if the contour were clockwise, then the answer would be $I = -(\pi e^2/3)i$. ▮

EXAMPLE 2. Evaluate

$$I = \oint_C \frac{\cos z}{(z+2)(z+i)(z-2i)}\, dz, \tag{7}$$

where C is as shown in Fig. 3a.

First, examine the integrand. It is singular at $z = -2, -i$, and $2i$. Of these three singular points, the latter two lie within C. Let us deform C into the "dumbell" contour $C_1 + C_2 + C_3 + C_4$ shown in Fig. 3b, where C_2 and C_4 are actually coincident (lying on the y axis) but are shown as slightly separated just for graphical clarity; such deformation is permissible, according to Theorem 23.3.2, because the integrand is analytic between and on C and the dumbell contour. Thus, using shorthand notation,

$$I = \oint_C = \oint_{C_1+C_2+C_3+C_4} = \oint_{C_1} + \int_{C_2} + \oint_{C_3} + \int_{C_4} = \oint_{C_1} + \oint_{C_3}, \tag{8}$$

where the last step follows from the fact that the integrals on C_2 and C_4 are negatives of each other. That is,

$$\begin{aligned} I &= \oint_{C_1} \left(\frac{\cos z}{(z+2)(z-2i)} \right) \frac{dz}{z+i} + \oint_{C_3} \left(\frac{\cos z}{(z+2)(z+i)} \right) \frac{dz}{z-2i} \\ &\equiv \oint_{C_1} \frac{f_1(z)}{z-a_1}\, dz + \oint_{C_3} \frac{f_3(z)}{z-a_3}\, dz. \end{aligned} \tag{9}$$

Now, the Cauchy integral formula (4) can be used to evaluate each of these two integrals. In the first, $f_1(z) = \cos z/[(z+2)(z-2i)]$ is analytic within C_1 and a_1 is $-i$, and in the second, $f_3(z) = \cos z/[(z+2)(z+i)]$ is analytic within C_3 and a_3 is $2i$ so (4) gives

$$\begin{aligned} I &= 2\pi i f_1(-i) + 2\pi i f_3(2i) \\ &= 2\pi i \left[\frac{\cos(-i)}{(2-i)(-3i)} + \frac{\cos(2i)}{(2+2i)(3i)} \right] \\ &= \frac{\pi}{30} \left[(5\cosh 2 - 8\cosh 1) - (5\cosh 2 + 4\cosh 1)i \right]. \end{aligned} \tag{10}$$

Alternatively, let us return to (5) and use partial fractions to express

$$\frac{1}{(z+i)(z-2i)} = -\frac{1}{3i}\frac{1}{z+i} + \frac{1}{3i}\frac{1}{z-2i}. \tag{11}$$

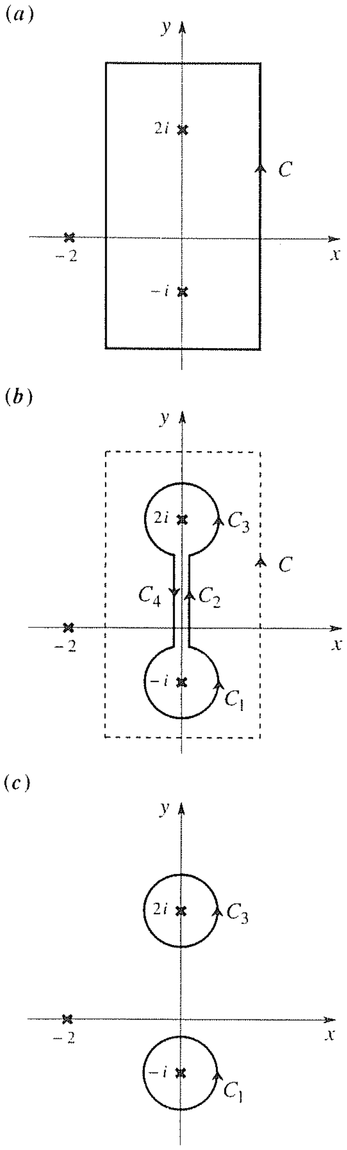


Figure 3. Divide and conquer.

Then (5) becomes

$$I = -\frac{1}{3i}\oint_C \left(\frac{\cos z}{z+2}\right)\frac{dz}{z+i} + \frac{1}{3i}\oint_C \left(\frac{\cos z}{z+2}\right)\frac{dz}{z-2i}. \tag{12}$$

In each integral in (12), $\cos z/(z+2)$ is analytic within C so the Cauchy integral formula gives

$$I = -\frac{2\pi i}{3i}\left(\frac{\cos z}{z+2}\right)\Big|_{z=-i} + \frac{2\pi i}{3i}\left(\frac{\cos z}{z+2}\right)\Big|_{z=2i}, \tag{13}$$

which reduces to the same answer as was given in (10).

COMMENT. In this example there were two singular points within C, whereas the Cauchy integral formula (4) allows only one. However, by deforming C into C_1 plus C_3 we were able to express I, in (9), as the sum of two integrals, each with only one singular point. Alternatively, that same objective can be met, without deforming C, by means of the partial fraction expansion (11) because each integrand in (12) has only one singular point within C. ▮

The Cauchy integral formula enables us to evaluate any integral, the integrand of which has a "first-order singularity" at some point a within the contour; it does not apply if the singularity is of second order or higher – that is, if I is of the form $\oint_C f(z)\,dz/(z-a)^2$, $\oint_C f(z)\,dz/(z-a)^3$, and so on. To deal with these cases, we begin by pointing out that (4) holds for *any* point a within C. If we emphasize the allowed variability of a by using the letter z in its place, and adopt a dummy integration variable ζ in place of z, then we can re-express (4) as

$$\oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta = 2\pi i f(z). \tag{14}$$

If we differentiate (14) with respect to z, and can justify the step

$$\frac{d}{dz}\oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta = \oint_C \frac{\partial}{\partial z}\left(\frac{f(\zeta)}{\zeta - z}\right)d\zeta = \oint_C \frac{f(\zeta)}{(\zeta - z)^2}\,d\zeta, \tag{15}$$

then we obtain

$$\oint_C \frac{f(\zeta)}{(\zeta - z)^2}\,d\zeta = 2\pi i f'(z) \tag{16}$$

for the evaluation of an integral with a "second-order singularity" in its integrand.

The first step in (15) does indeed need justification because it amounts to an interchange in the order of the two limit processes: the integration and the differentiation. The step *looks* all right because z is inside of C, so that $\zeta - z$ is nonzero for all ζ on C. Thus, $f(\zeta)/(\zeta - z)$ is an analytic function of z for each ζ on C. Nevertheless, let us verify (15) rigorously. Not having a complex variable version of the Leibniz rule for differentiating under the integral sign, let us fall back on basics and recall that the derivative is the limit of a difference quotient. Let

$$I(z) = \oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta. \tag{17}$$

Then

$$\begin{aligned}I'(z) &= \lim_{\Delta z\to 0} \frac{I(z+\Delta z)-I(z)}{\Delta z}\\ &= \lim_{\Delta z\to 0} \frac{1}{\Delta z}\oint_C \left(\frac{1}{\zeta - z - \Delta z} - \frac{1}{\zeta - z}\right) f(\zeta)\,d\zeta\\ &= \lim_{\Delta z\to 0} \oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z)(\zeta - z - \Delta z)}.\end{aligned} \tag{18}$$

If we pass the limit across the integral sign, then the latter does give the same result as (15) but, as noted above, we cannot be sure that the order of the limit and the integration can be reversed. To proceed, we express

$$\oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z)(\zeta - z - \Delta z)} = \oint_C \frac{f(\zeta)}{(\zeta - z)^2}\,d\zeta + \Delta I \tag{19}$$

and seek to show that $\Delta I \to 0$ as $\Delta z \to 0$. From (19), ΔI is the difference

$$\begin{aligned}\Delta I &= \oint_C \left[\frac{1}{(\zeta - z)(\zeta - z - \Delta z)} - \frac{1}{(\zeta - z)^2}\right] f(\zeta)\,d\zeta\\ &= \Delta z \oint_C \frac{f(\zeta)\,d\zeta}{(\zeta - z)^2(\zeta - z - \Delta z)}.\end{aligned} \tag{20}$$

To bound ΔI let us deform C in (20) to a circle C' with its center at z and radius ρ small enough so that C' lies entirely inside C. Since we are letting $\Delta z \to 0$, we can choose $|\Delta z| < \rho/5$, say. Then, from Fig. 4 we can see that $|(\zeta - z)^2|$ is exactly equal to ρ^2 for all points ζ on C', and that $|\zeta - (z + \Delta z)| > 4\rho/5$ for all ζ on C'. Moreover, there must exist some finite constant m such that $|f(\zeta)| \le m$ on C' since f is analytic on C'. Therefore, the ML bound gives

$$|\Delta I| \le |\Delta z| \frac{m}{\rho^2(4\rho/5)} 2\pi\rho. \tag{21}$$

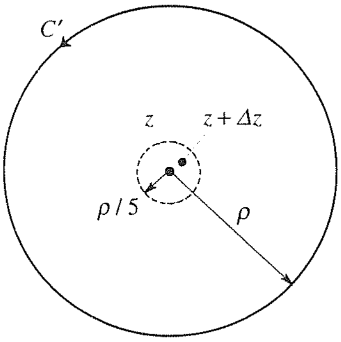


Figure 4. Bounding ΔI.

It follows from (21) that if $\Delta z \to 0$ with C' fixed (and hence with m and ρ fixed), then $\Delta I \to 0$, which result establishes the truth of (15) and hence (16).

Similarly, we can repeat this process of differentiation as many times as we wish, and find that

$$\oint_C \frac{f(\zeta)}{(\zeta - z)^{n+1}}\,d\zeta = \frac{2\pi i}{n!} f^{(n)}(z)$$

or, returning to the z, a notation used in (4),

$$\boxed{\oint_C \frac{f(z)}{(z-a)^{n+1}}\,dz = \frac{2\pi i}{n!} f^{(n)}(a).} \tag{22}$$

For reference, let us distinguish (4) and (22) as the Cauchy integral formula and the **generalized Cauchy integral formula**, respectively.*

EXAMPLE 3. Evaluate

$$I = \oint_C \frac{e^z}{z^3}\,dz, \tag{23}$$

where C is the counterclockwise unit circle $|z| = 1$. Rewrite (23) in the form

$$I = \oint_C \frac{e^z}{(z-0)^3}\,dz, \tag{24}$$

for comparison with (22). We see that $n = 2$, $a = 0$, and $f(z) = e^z$ so (22) gives

$$I = \frac{2\pi i}{2!}\left(\frac{d^2}{dz^2}e^z\right)\bigg|_{z=0} = \pi i. \quad ∎$$

EXAMPLE 4. Evaluate

$$I = \oint_C \frac{z+1}{z(z-2)(z-4)^3}\,dz, \tag{25}$$

where C is the counterclockwise unit circle $|z-3| = 2$ (Fig. 5a). The integrand has singularities at $z = 0, 2,$ and 4, of which the latter two fall within C. We could apply partial fractions to the $1/[(z-2)(z-4)^3]$ part of the integrand, but it is easier to deform C into the two contours shown in Fig. 5b, and to evaluate each of the two integrals using (22). Thus,

$$\begin{aligned} I &= \oint_{C_1}\left[\frac{z+1}{z(z-4)^3}\right]\frac{dz}{z-2} + \oint_{C_2}\left[\frac{z+1}{z(z-2)}\right]\frac{dz}{(z-4)^3} \\ &= 2\pi i\left[\frac{z+1}{z(z-4)^3}\right]\bigg|_{z=2} + \frac{2\pi i}{2!}\frac{d^2}{dz^2}\left[\frac{z+1}{z(z-2)}\right]\bigg|_{z=4} \\ &= -\frac{3\pi i}{8} + \frac{23\pi i}{64} = -\frac{\pi i}{64}. \quad ∎ \end{aligned}$$

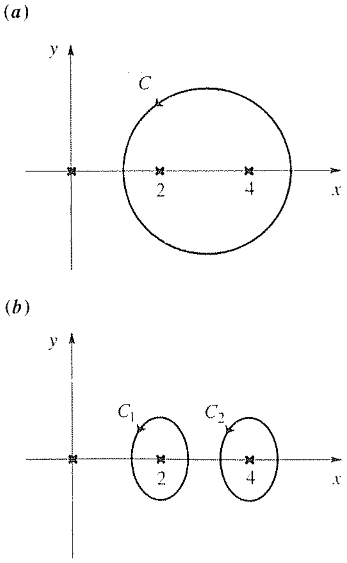


Figure 5. The contours in Example 4.

There are two important observations to be made. First, observe that the Cauchy integral formula

$$f(z) = \frac{1}{2\pi i}\oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta \tag{26}$$

gives $f(z)$ at any interior point z of the region within C as an integration over its boundary values on C. Since the real and imaginary parts of the analytic function

*This result opens up the interesting possibility of defining noninteger-order derivatives of a function, for if we interpret $n!$ as $\Gamma(n+1)$ and define the intended branch of $(\zeta - z)^{n+1}$, then the right-hand side of (22) is defined for noninteger values of n as well.

$f(z) = u(x, y) + iv(x, y)$ are harmonic, we see that (26) is very close to providing the solution to the classical Dirichlet problem, namely, $\nabla^2 u = 0$ in the interior of C together with a Dirichlet boundary condition on u (i.e., u given on C). Thus, as in Chapter 22, we see once again the close connection between analytic function theory and two-dimensional potential theory. We return to this idea in the exercises.

Second, observe that *having assumed only that $f(z)$ is analytic (once differentiable), one finds with no further assumption that $f(z)$ possesses derivatives of all orders*:

$$f^{(n)}(z) = \frac{n!}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z)^{n+1}}\, d\zeta. \tag{27}$$

This remarkable result has no analog in real variable theory. For instance, observe that $f(x) = x^2 H(x)$, where H is the Heaviside function, is once differentiable for all x (sketch the graphs of f and f'), but f'' fails to exist at $x = 0$ due to the kink in the graph of f' at that point. Similarly, $f(x) = x^3 H(x)$ is differentiable twice but not three times. Thus, if $f(z)$ is "nice enough" to be once differentiable in some region, then it will be very nice indeed, *infinitely* differentiable! How can we understand this difference between real and complex function theory? The answer is that differentiability in the complex plane is much more demanding than on the real axis because z can tend to z_0, in the formula

$$f'(z) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}, \tag{28}$$

in any manner, whereas only a horizontal approach is possible in the real variable formula

$$f'(x) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}. \tag{29}$$

Because more is demanded, in the complex case, the consequences of differentiability are more far reaching.

Closure. The chief results in this section are the Cauchy integral formula (4) and its generalized version (22). Whereas Cauchy's theorem tells us that $\oint_C f(z)\, dz = 0$ if f is analytic, (22) shows us how to evaluate an integral (around a closed contour C) when its integrand has a singularity of any order within the region enclosed by C.

EXERCISES 23.5

1. Evaluate each integral, where C is the counterclockwise circle $|z| = 3$. Use Cauchy's integral formula or its extension.

(a) $\displaystyle\oint_C \frac{\cos z}{z}\, dz$

(b) $\displaystyle\oint_C \frac{\sin z}{z}\, dz$

(c) $\displaystyle\oint_C \frac{dz}{z^2 - 5z}$

(d) $\displaystyle\oint_C \frac{z^2 - 1}{z^2 + 1}\, e^z\, dz$

(e) $\displaystyle\oint_C \frac{z + 1}{(z - 1)(z + 2)^3}\, dz$

(f) $\displaystyle\oint_C \frac{e^{2z}}{z^5}\, dz$

(g) $\displaystyle\oint_C \frac{\sinh 3z}{(z^2+1)^2}\,dz$ (h) $\displaystyle\oint_C \frac{z+2}{z^4-1}\,dz$

(i) $\displaystyle\oint_C \frac{e^{z^2}}{z\cos(z/2)}\,dz$ (j) $\displaystyle\oint_C \frac{z\,dz}{(z+i)(z^2+1)}$

(k) $\displaystyle\oint_C \frac{z^3}{z^2+i}\,dz$ (l) $\displaystyle\oint_C \frac{dz}{z(z-2)(z-4)}$

2. (*Important little integral*) In Section 23.3 we show that

$$\oint_C (z-a)^n\,dz = \begin{cases} 2\pi i & (n=-1) \\ 0 & (n \neq -1), \end{cases} \tag{2.1}$$

where n is any integer and a is within the contour C. Derive (2.1) using Cauchy's theorem, the Cauchy integral formula, and the generalized Cauchy integral formula.

3. (a) Show from (22) that if C is a circle of radius ρ with center at z, $f(z)$ is analytic inside and on C, and M is the maximum value of $|f(z)|$ on C, then

$$\left|f^{(n)}(z)\right| \le \frac{n!\,M}{\rho^n}. \tag{3.1}$$

(b) (*Liouville's theorem*) Use (3.1) to prove **Liouville's theorem**: *If f is entire (i.e., analytic for all finite z) and bounded for all z, then f is a constant.*

(c) Since $f(z) = \sin z$ is entire and not a constant, it must not be bounded (according to Liouville's theorem). Demonstrate that, in fact, it is *not* bounded.

(d) (*Fundamental theorem of algebra*) Use Liouville's theorem to prove the **fundamental theorem of algebra**: *if $P(z)$ is a polynomial function of z, of degree 1 or greater,*

$$P(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0 \qquad (a_n \neq 0)$$

then $P(z) = 0$ has at least one root. HINT: Suppose that $P(z)$ is nonzero everywhere. Then $f(z) = 1/P(z)$ is analytic everywhere and is bounded.

4. (*Dirichlet problems*) As mentioned in the text, just as the Cauchy integral formula

$$f(z) = \frac{1}{2\pi i}\oint_C \frac{f(\zeta)}{\zeta - z}\,d\zeta \tag{4.1}$$

expresses an analytic function $f(z) = u + iv$ in terms of its boundary values, we would expect there to exist a similar integral formula expressing a harmonic function $u(x, y)$ in terms of its boundary values. In this exercise we seek such a formula for two important cases: the case where the domain is a circular disk, and the case where the domain is the upper half plane.

(a) (*Circular disk*) Let C be the counterclockwise circle $|\zeta| = R$. If we seek the desired expression for u by equating real parts of the left- and right-hand sides of (4.1), we find that the right-hand side involves both u *and* v, whereas the additional unknown v is not welcome. The reason that v enters is that $1/(\zeta - z)$ is not purely real. With $\zeta = Re^{i\phi}$, show that we can re-express (4.1) as

$$\begin{aligned} f(z) &= \frac{1}{2\pi i}\oint_C \left(\frac{1}{\zeta - z} - \frac{1}{\zeta - R^2/\overline{z}}\right) f(\zeta)\,d\zeta \\ &= \frac{1}{2\pi}\int_0^{2\pi} \left(\frac{\zeta}{\zeta - z} + \frac{\overline{z}}{\overline{\zeta} - \overline{z}}\right) f(\zeta)\,d\phi, \end{aligned} \tag{4.2}$$

where the bracketed quantity is real. In particular, show that

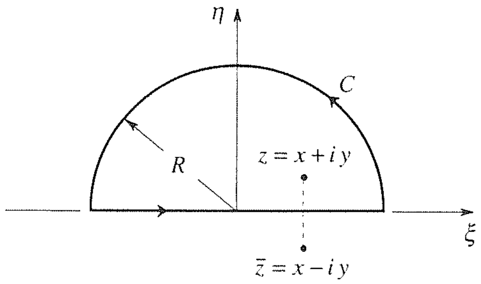
$$\frac{\zeta}{\zeta - z} + \frac{\overline{z}}{\overline{\zeta} - \overline{z}} = \frac{R^2 - r^2}{|\zeta - z|^2}, \tag{4.3}$$

and hence that

$$u(r, \theta) = \frac{1}{2\pi}\int_0^{2\pi} \frac{(R^2 - r^2)\,u(R, \phi)}{R^2 - 2Rr\cos(\phi - \theta) + r^2}\,d\phi, \tag{4.4}$$

where $z = re^{i\theta}$ and $\zeta = Re^{i\phi}$. This result is also derived by separation of variables, in Section 20.3 and is known as the **Poisson integral formula for the circular disk**.

(b) (*Upper half plane*) This time let C be the contour shown here. Show that (4.1) can be re-expressed as



$$f(z) = \frac{1}{2\pi i}\oint_C \left(\frac{1}{\zeta - z} - \frac{1}{\zeta - \overline{z}}\right) f(\zeta)\,d\zeta \tag{4.5}$$

for all $R > |z|$. Suppose that, as our boundary condition at infinity, $f(z) \to 0$ as $z \to \infty$. Letting $R \to \infty$ in (4.5), show that the semicircle integral tends to zero, leaving us with

$$f(z) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \left(\frac{1}{\xi - z} - \frac{1}{\xi - \overline{z}}\right) f(\xi)\,d\xi. \tag{4.6}$$

Finally, equating real parts in (4.6), show that

$$u(x,y) = \frac{y}{\pi}\int_{-\infty}^{\infty} \frac{u(\xi,0)}{(\xi - x)^2 + y^2}\,d\xi \qquad (4.7)$$

is the solution to the Dirichlet problem for the upper half plane, with the boundary condition $u(x,y) \to 0$ as $r = \sqrt{x^2+y^2} \to \infty$. This result is also derived by means of the Fourier transform in Section 20.4 and is the **Poisson integral formula for the upper half plane**. NOTE: The Dirichlet problems considered here in parts (a) and (b) can also be solved by an elegant technique known as the method of **Green's functions**.* For the reader who subsequently studies that method and comes back to this exercise to compare the two approaches, it may be helpful for us to note that the mysterious point $\zeta = R^2/\overline{z}$ that shows up in (4.2) and the mysterious point $\zeta = \overline{z}$ that shows up in (4.5) are (in the terminology of the method of Green's functions) the *image* points.

Chapter 23 Review

We begin Chapter 23 by defining the complex integral $\int_C f(z)\,dz$ in essentially the same way that one defines the real integral $\int_a^b f(x)\,dx$, and note the close resemblance to line integrals in two dimensions. There are only a handful of major results, and they are as follows.

Cauchy's Theorem. If $f(z)$ is analytic in a simply connected domain D, then

$$\oint_C f(z)\,dz = 0$$

for every piecewise smooth simple closed curve C in D.

If instead C is an open curve, then Cauchy's theorem gives this corollary:

Path Independence. If $f(z)$ is analytic in a simply-connected domain D, then $\int_C f(z)\,dz$ is independent of path in D. That is, given any initial point P in D and any final point Q in D, the value of $\int_C f(z)\,dz$ is the same for every piecewise smooth path C, lying entirely within D, from P to Q.

Analogous to the fundamental theorem of the real integral calculus, we have:

Fundamental Theorem of the Complex Integral Calculus. Let $f(z)$ be analytic in a simply-connected domain D, and let z_0 be any fixed point in D. Then

(i) $G(z) = \int_{z_0}^{z} f(\zeta)\,d\zeta$ is analytic in D and $G'(z) = f(z)$.

(ii) If $F(z)$ is any primitive of $f(z)$ [i.e., $F'(z) = f(z)$], then

$$\int_{z_0}^{z} f(\zeta)\,d\zeta = F(z) - F(z_0).$$

*See, for example, M. D. Greenberg, *Applications of Green's Functions in Science and Engineering* (Englewood Cliffs, NJ: Prentice Hall, 1971).

If the integrand is *not* analytic within the closed contour C, then we have:

Cauchy Integral Formula. Let $f(z)$ be analytic in a simply-connected domain D, let C be a piecewise smooth simple closed counterclockwise curve in D, and let a be any fixed point lying within C. Then

$$\oint_C \frac{f(z)}{z-a}\,dz = 2\pi i f(a).$$

In fact, for any $n = 0, 1, 2, \ldots$ we have the generalized Cauchy integral formula

$$\oint_C \frac{f(z)}{(z-a)^{n+1}}\,dz = \frac{2\pi i}{n!} f^{(n)}(a).$$

Note that we have *not* claimed that the latter formula covers all possible singular integrands, but we do claim that it suffices for most applications, and is of great importance.

Finally, we also develop the so-called ML bound:

$$\left| \int_C f(z)\,dz \right| \leq ML,$$

if $|f(z)| \leq M$ on C and the length of C is L. We continue to use this bound in the next chapter, both as a theoretical tool in the derivatives, and in the applications as well.

Chapter 24

Taylor Series, Laurent Series, and the Residue Theorem

24.1 Introduction

When we first introduced complex integrals, in Section 23.2, we pointed out that we can express any complex integral $\int_C f(z)\,dz$ in the form

$$\int_C f(z)\,dz = \int_C [u(x,y)dx - v(x,y)dy] + i\int_C [v(x,y)dx + u(x,y)dy].$$

Hence, there was the temptation to stop right there and to rely on the theory of real line integrals, as is covered in Chapter 15 and 16. However, we promised that it would prove more fruitful to keep $\int_C f(z)\,dz$ intact and to develop a distinct complex integral calculus. We began that development in Chapter 23 and got as far as Cauchy's theorem, the fundamental theorem of the complex integral calculus, and the Cauchy integral formula. Those results now enable us to derive the Taylor and Laurent series expansions of functions of a complex variable which, in turn, permits us to clarify the notion of singularities and, finally, to complete our development of the complex integral calculus by deriving the powerful residue theorem.

24.2 Complex Series and Taylor Series

24.2.1. Complex series. By a **complex series** we mean any sum of the form

$$\sum_{n=1}^{\infty} c_n = c_1 + c_2 + c_3 + \cdots, \tag{1}$$

where the c_n's are complex numbers. As in the real case, we say that the series converges if the limit of the sequence of partial sums, $s_n = c_1 + c_2 + \cdots + c_n$, exists as $n \to \infty$. That is, the series **converges** to s if to each (real) number $\epsilon > 0$, no matter how small, there exists an integer $N(\epsilon)$ such that $|s_n - s| < \epsilon$ for all $n > N$; otherwise, it **diverges**.

The sum $\sum c_n$ can be expressed in terms of real series, for if $c_n = a_n + ib_n$ where a_n and b_n are real, then $\sum c_n$ converges if and only if $\sum a_n$ and $\sum b_n$ do, in which case $\sum c_n = \sum a_n + i \sum b_n$ (Exercise 1). However, just as it is best to develop a complex integral calculus that does not rely on expressing $\int_C f(z)\, dz$ in terms of real integrals, likewise it is best to develop a theory of complex series.

As a start, a necessary and sufficient condition for convergence of any series, real or complex, is given by the Cauchy convergence theorem:

THEOREM 24.2.1 *Cauchy Convergence Theorem*
An infinite series is convergent if and only if its sequence of partial sums s_n is a Cauchy sequence – that is, if to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| < \epsilon$ for all m and n greater than N.

Unfortunately, this theorem is difficult to apply so one develops (in the calculus) an array of theorems (i.e., tests for convergence/divergence), that are more specialized than the Cauchy convergence theorem, but easier to apply. For instance, if in Theorem 24.2.1 we set $m = n - 1$, then the stated condition becomes: to each $\epsilon > 0$ (no matter how small) there corresponds an integer $N(\epsilon)$ such that $|s_m - s_n| = |c_n| < \epsilon$ for all $n > N$, which is equivalent to saying that $c_n \to 0$ as $n \to \infty$. Thus, we have the following specialized, but readily applied, result:

THEOREM 24.2.2 *A Necessary Condition for Convergence*
For the series $\sum_{n=1}^{\infty} c_n$ to converge it is necessary, but not sufficient, that $c_n \to 0$ as $n \to \infty$.

Two more such theorems follow.

THEOREM 24.2.3 *Comparison Test*
If the series $\sum_{n=1}^{\infty} M_n$ converges, where the M_n's are positive constants and $|c_n| \leq M_n$ for each n greater than some integer N, then the series $\sum_{n=1}^{\infty} c_n$ converges too.

Proof: According to the Cauchy criterion, the assumed convergence of $\sum_{n=1}^{\infty} M_n$ implies that to each $\epsilon > 0$ there corresponds an N_0 such that $M_p + M_{p+1} + \cdots + M_{p+q} < \epsilon$ for all $p > N_0$ and all $q \geq 0$. Then (Exercise 2),

$$\begin{aligned} |c_p + c_{p+1} + \cdots + c_{p+q}| &\leq |c_p| + |c_{p+1}| + \cdots + |c_{p+q}| \\ &\leq M_p + M_{p+1} + \cdots + M_{p+q} < \epsilon \end{aligned} \tag{2}$$

for all $p > N_0$ (if $N_0 < N$, then use N instead) and for all $q \geq 0$, so $\sum_{n=1}^{\infty} c_n$ converges too. ■

EXAMPLE 1. Determine the convergence or divergence of the series

$$\sum_{n=0}^{\infty} \frac{i^n}{n!} = 1 + i + \frac{i^2}{2!} + \frac{i^3}{3!} + \cdots . \tag{3}$$

Since

$$\left| \frac{i^n}{n!} \right| = \frac{1}{n!} < \frac{1}{2^n} \tag{4}$$

for all $n \geq 4$, and $\sum_{n=0}^{\infty} 1/2^n = \sum_{n=0}^{\infty} (\frac{1}{2})^n$ is a convergent geometric series, it follows from the comparison test that (3) is convergent. ■

Observe that sometimes our series start at $n = 0$ and sometimes at $n = 1$ or some other integer. These differences are inconsequential insofar as convergence/divergence are concerned (assuming that each of the initial terms in question is finite) though they do of course affect the sum if the series is convergent.

THEOREM 24.2.4 *Ratio Test*
If $\lim_{n\to\infty} |c_{n+1}/c_n| = L$, then $\sum_{n=0}^{\infty} c_n$ converges if $L < 1$ and diverges if $L > 1$. No information is obtained if $L = 1$ or if the limit does not exist.

Proof: Suppose that $L < 1$. Choose any number p such that $L < p < 1$. Then it must be true that $|c_{n+1}/c_n| < p$ for all sufficiently large n's, say for $n \geq N$. Thus,

$$\begin{aligned} &|c_{N+1}| < p\,|c_N|, \\ &|c_{N+2}| < p\,|c_{N+1}| < p^2\,|c_N|, \\ &|c_{N+3}| < p\,|c_{N+2}| < p^3\,|c_N|, \end{aligned}$$

and so on. According to the comparison test, $\sum_{n=0}^{\infty} c_n$ converges by comparison with the series $\sum_{n=0}^{\infty} p^n\,|c_N|$, which is $|c_N|$ times the convergent geometric series $\sum_{n=0}^{\infty} p^n$, convergent because $p < 1$. The case $L > 1$ is left for the exercises. ■

EXAMPLE 2. Determine the convergence or divergence of the series

$$\sum_{n=0}^{\infty} \frac{(1+i)^n}{n!}. \tag{5}$$

In this case,

$$\lim_{n\to\infty} \left| \frac{c_{n+1}}{c_n} \right| = \lim_{n\to\infty} \left| \frac{1+i}{n+1} \right| = \sqrt{2} \lim_{n\to\infty} \frac{1}{n+1} = 0 \tag{6}$$

so $L = 0$ and, according to the ratio test, the series (5) is convergent. ▮

EXAMPLE 3. Determine the convergence or divergence of the series

$$\sum_{n=2}^{\infty} e^{-(2+3i)n}. \tag{7}$$

Again applying the ratio test,

$$\begin{aligned} \lim_{n\to\infty}\left|\frac{c_{n+1}}{c_n}\right| &= \lim_{n\to\infty}\left|e^{-(2+3i)}\right| = \left|e^{-(2+3i)}\right| \\ &= \left|e^{-2}e^{-3i}\right| = e^{-2}\left|e^{-3i}\right| = e^{-2} \end{aligned} \tag{8}$$

so $L = e^{-2} < 1$ and the series (7) is convergent. ▮

More generally, the terms may be functions of z rather than constants:

$$\sum_{n=0}^{\infty} f_n(z) = f_0(z) + f_1(z) + \cdots. \tag{9}$$

Then the set of all points in the z plane for which the series converges is called the *region of convergence* of the series.

Since we are leading up to complex Taylor series we shall not consider the general case (9). Rather, we suppose that $f_n(z)$ is of the form $c_n(z-a)^n$, where the c_n's and a are, in general, complex numbers. The resulting form,

$$\boxed{\sum_{n=0}^{\infty} a_n(z-a)^n = a_0 + a_1(z-a) + a_2(z-a)^2 + \cdots,} \tag{10}$$

is called a **power series**. If, to determine the region of convergence of (10), we apply the ratio test, we obtain the following result, proof of which is left for the exercises.

THEOREM 24.2.5 *Power Series Convergence by Ratio Test*
If

$$\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = L, \tag{11}$$

then the power series (10) converges in the disk $|z-a| < 1/L$ and diverges in $|z-a| > 1/L$. If $L = \infty$ the series converges only at the point $z = a$, and if $L = 0$ it converges for all z, that is, in $|z-a| < \infty$.

EXAMPLE 4. *Geometric Series.* For the **geometric series**

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots, \tag{12}$$

$a_n = 1$ for each n so (11) gives $L = 1$. Hence, the geometric series (12) converges in the disk $|z| < 1$ and diverges in $|z| > 1$. Theorem 24.2.5 gives no information regarding convergence or divergence for points on the circele $|z| = 1$. However, $|z^n| = |z|^n = 1^n = 1$ fails to tend to zero as $n \to \infty$ so, by Theorem 24.2.2, the series diverges at all points on the circle $|z| = 1$. ▮

It is tempting to conclude from Theorem 24.2.5 that a power series inevitably converges in a *circular disk* centered at a (that disk being the single point $z = a$ if $L = \infty$ and the whole z plane if $L = 0$). However, notice that the theorem yields no information if $\lim_{n\to\infty} |a_{n+1}/a_n|$ does not exist (as occurs if, for example, $a_n = \sin n$), and in that event it is conceivable that a noncircular region of convergence may exist. Toward clarifying this point, we first present the following lemma.*

LEMMA 24.2.1 *Power Series Convergence*
If the power series (10) converges at z_0, then it converges everywhere in the open disk $|z - a| < |z_0 - a|$ (Fig. 1).

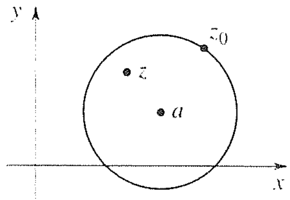


Figure 1. The disk $|z - a| < |z_0 - a|$.

Proof: Since the series converges for $z = z_0$, it follows that $a_n(z_0 - a)^n \to 0$ as $n \to \infty$. Thus, there must certainly exist a constant K such that $|a_n(z_0 - a)^n| < K$ for all $n = 0, 1, 2, \ldots$, and from this result we infer that $|a_n| < K/|z_0 - a|^n$ $(n = 0, 1, 2, \ldots)$. Then, for any z such that $|z - a| < |z_0 - a|$, we have

$$|a_n(z_0 - a)^n| < K\left|\frac{z - a}{z_0 - a}\right|^n. \tag{13}$$

Observing that $|(z - a)/(z_0 - a)| < 1$, we see that $K|(z - a)/(z_0 - a)|^n$ is the nth term of a convergent geometric series (scaled by K). Thus, it follows from (13) and the comparison test that (10) is convergent, as stated. ■

With the help of this lemma we can now show that the region of convergence of a power series is necessarily a circular disk.

THEOREM 24.2.6 *Power Series Convergence in a Disk*
For the power series (10) there exist three possibilities regarding its region of convergence:

*Not an end in itself, a lemma is developed to help prove a subsequent theorem.

(i) the series converges only at the point a;
(ii) the series converges in the whole plane;
(iii) there exists a constant $R > 0$, called the **radius of convergence** of (10), such that (10) converges in $|z - a| < R$ and diverges in $|z - a| > R$.

Proof: To prove that (i) and (ii) are possibilities it suffices to put forward an example of each case. We leave this part for the exercises, and turn to (iii). If (i) does *not* apply, there must be a point $z_1 \neq a$ at which (10) converges, and if (ii) does *not* apply, there must be a point z_2 at which (10) diverges. From the lemma it is evident that $|z_2 - a| \geq |z_1 - a|$. If $|z_2 - a| = |z_1 - a|$, it follows from the lemma that (10) converges everywhere in $|z - a| < |z_1 - a|$. Furthermore, the lemma also implies that (10) diverges everywhere in $|z - a| > |z_1 - a|$ for if it converged anywhere in that region it could not diverge at z_2 as assumed. Thus, if $|z_2 - a| = |z_1 - a|$, then $R = |z_2 - a| = |z_1 - a|$ is the radius of convergence referred to in (iii), and it remains to consider the case where $|z_2 - a| > |z_1 - a|$ as depicted in Fig. 2. Let S be the set of values ρ such that (10) converges for all $|z - a| < \rho$. From the lemma it follows that if some value ρ_0 belongs to S so do all ρ's satisfying $0 \leq \rho < \rho_0$. Thus, S is necessarily an interval (on a ρ axis), either $0 \leq \rho < R$ or $0 \leq \rho \leq R$ for some R such that $|z_1 - a| \leq R \leq |z_2 - a|$. Furthermore, (10) must diverge everywhere in $|z - a| > R$ because if it converged anywhere in that region, the right endpoint of the S interval would be greater than R. Then R is the radius of convergence referred to in (iii), and the proof is complete. ■

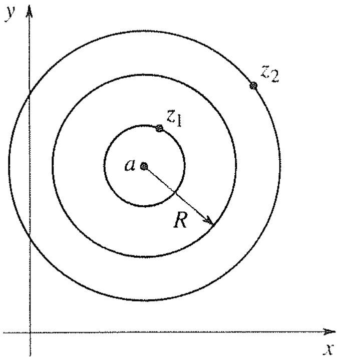


Figure 2. Disk of convergence.

We are now ready to develop the concept of the Taylor series of a function of a complex variable.

24.2.2. Taylor series. In Section 23.5 we found that if a function $f(z)$ is analytic at $z = a$, then it admits derivatives of all orders there. Since $f(a), f'(a), f''(a), \ldots$ all exist, it is at least formally possible to write down the **Taylor series**

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(z - a)^n \tag{14}$$

of $f(z)$ about the point a. However, whether or not the Taylor series is *useful* depends upon the answers to these questions: Does the series converge and, if so, in what domain? If it does converge in some domain D, does it converge to $f(z)$ there; that is, does it *represent* f in D? Finally, if (14) does represent f in D, might there exist *other* power series representations of f as well? That is, might there exist more than one representation of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - a)^n,$$

where $a_n \neq f^{(n)}(a)/n!$? These questions are answered by the following theorem.

THEOREM 24.2.7 *Taylor Series*
Let $f(z)$ be analytic in a domain D. If a lies within D and $|z-a|<R$ is any disk centered at a and lying entirely within D, then $f(z)$ admits precisely one power series representation in $|z-a|<R$, its Taylor series (14):

$$\boxed{f(z)=\sum_{n=0}^{\infty}\frac{f^{(n)}(a)}{n!}(z-a)^n=f(a)+f'(a)(z-a)+\cdots.} \tag{15}$$

Proof: Let C be the circle $|z-a|=R$ (Fig. 3), counterclockwise. For any point z inside C we have, according to the Cauchy integral formula,

$$f(z)=\frac{1}{2\pi i}\oint_C\frac{f(\zeta)}{\zeta-z}\,d\zeta, \tag{16}$$

which formula serves as our starting point. Let us re-express

$$\frac{1}{\zeta-z}=\frac{1}{\zeta-a}\,\frac{1}{1-\dfrac{z-a}{\zeta-a}} \tag{17}$$

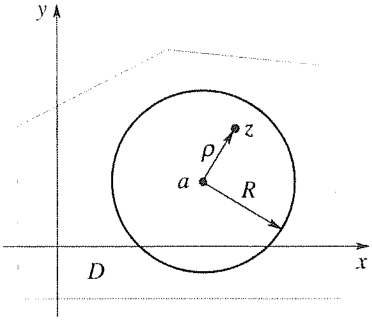


Figure 3. The contour C in (16).

in the integrand, and then use the formula

$$\frac{1}{1-t}=1+t+t^2+\cdots+t^{n-1}+\frac{t^n}{1-t}, \tag{18}$$

which is simply an algebraic identity for all $t\neq 1$ [as can be verified by multiplying (18) through by $1-t$]. Specifically, let $(z-a)/(\zeta-a)$ be t in (17), and then use (18), These steps give

$$\begin{aligned} f(z)=\frac{1}{2\pi i}\Bigg[\oint_C\frac{f(\zeta)}{\zeta-a}\,d\zeta+(z-a)\oint_C\frac{f(\zeta)}{(\zeta-a)^2}\,d\zeta+\cdots \\ +(z-a)^{n-1}\oint_C\frac{f(\zeta)}{(\zeta-a)^n}\,d\zeta\Bigg]+R_n(z), \end{aligned} \tag{19}$$

where the *remainder term* is*

$$R_n(z)=\frac{(z-a)^n}{2\pi i}\oint_C\frac{f(\zeta)}{(\zeta-a)^n(\zeta-z)}\,d\zeta. \tag{20}$$

*At first glance, it may appear that in deriving (19) we have been guilty of interchanging two limit processes without justification, namely, expressing the integral of a sum as the sum of the integrals. However, there is no need for concern because $1+t+t^2+\cdots+t^{n-1}$ is not an infinite series; it is only a finite sum.

Then, recalling the generalized Cauchy integral formula

$$f^{(n)}(a) = \frac{n!}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - a)^{n+1}}\, d\zeta, \tag{21}$$

we see that (19) becomes

$$f(z) = f(a) + f'(a)(z-a) + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(z-a)^{n-1} + R_n, \tag{22}$$

which is the complex version of Taylor's formula with remainder.

Finally, we wish to show that $R_n \to 0$ as $n \to \infty$. We expect that to be true since R_n comes from the $t^n/(1-t)$ term in (18), and t^n does tend to zero as $n \to \infty$ because $|t| = |(z-a)/(\zeta - a)| = \rho/R < 1$ for all points ζ on C (Fig. 3).

Since $f(z)$ is analytic on C it must be bounded on C. Thus, suppose that $|f(z)| < m$ on C. Noting further, from Fig. 3, that $|z-a| = \rho$, that $|\zeta - a| = R$, and that $|\zeta - z| \geq R - \rho$ for all ζ on C, we have

$$\begin{aligned} |R_n(z)| &= \frac{\rho^n}{2\pi}\left| \oint_C \frac{f(\zeta)}{(\zeta-a)^n(\zeta-z)}\, d\zeta \right| \\ &\leq \frac{\rho^n}{2\pi}\frac{m}{R^n(R-\rho)} 2\pi R \qquad \text{(by the } ML \text{ bound)} \\ &= \frac{mR}{R-\rho}\left(\frac{\rho}{R}\right)^n, \end{aligned} \tag{23}$$

which expression tends to zero as $n \to \infty$ for all z in $|z-a| < R$. Then, by the definition of series convergence, letting $n \to \infty$ in (23) gives the resulting Taylor series representation

$$f(z) = f(a) + f'(a)(z-a) + \frac{f''(a)}{2!}(z-a)^2 + \cdots \tag{24}$$

of $f(z)$ in $|z-a| < R$.

We will omit the proof of uniqueness, that (24) is the only representation of $f(z)$ of the form $\sum_{n=0}^{\infty} a_n (z-a)^n$. ■

The upshot is that if $f(z)$ is analytic in $|z-a| < R$, then it can be represented in that disk by a convergent power series, namely, its Taylor series about a. It can also be shown that the converse is true:

THEOREM 24.2.8 *Convergent Power Series*
If a power series $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges in $|z-a| < R$, then its sum function $f(z)$ is analytic there and the power series is the Taylor series of f; that is, $a_n = f^{(n)}(a)/n!$.

EXAMPLE 5. To obtain the Taylor series expansion of $f(z) = e^z$ about $z = 0$, say, we could work out the coefficients $f(0), f'(0), f''(0)/2!$, and so on, but it is easier to note that (15) is of exactly the same form as in real variable theory so we need merely change the x's to z's in the familiar formula

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots. \tag{25}$$

Observing also that e^z is analytic in the whole plane, we see from Theorem 24.2.7 that the Taylor series representation

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots \tag{26}$$

holds in $|z| < \infty$. As in real variable theory, we call the Taylor expansion of $f(z)$ about $z = 0$ the Maclaurin expansion of $f(z)$.

If, however, we wish to expand e^z about some point other than the origin, say $z = i$, then the remembered formula (25) is of no help, and we need to work out the $f^{(n)}(a)/n!$ coefficients in (15). Doing so, we obtain

$$e^z = e^i + e^i(z-i) + \frac{e^i}{2!}(z-i)^2 + \cdots, \tag{27}$$

which holds in the whole plane, $|z - i| < \infty$. Of course, $e^i = \cos 1 + i \sin 1$. ▮

EXAMPLE 6. *Geometric Series Revisited.* In Example 4 we used Theorem 24.2.5 to determine that the geometric series $1 + z + z^2 + \cdots$ converges in $|z| < 1$ and diverges in $|z| > 1$. (We also showed that it diverges on the circle $|z| = 1$ because there the general term z^n fails to go to zero as $n \to \infty$.) In fact, the series sums to $1/(1-z)$ in $|z| < 1$ and is the Taylor series of that function:

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots. \qquad (|z| < 1) \tag{28}$$

It is important to understand that the preceding three theorems enable us to determine the region of convergence directly from the function being expanded, without having to test the series itself. To illustrate, consider $f(z) = 1/(1-z)$. Theorem 24.2.6 tells us that the region of convergence of the Taylor series of f about $z = 0$ will be a disk centered at $z = 0$, having zero, finite, or infinite radius. To determine that radius, examine f to see where it is analytic and where it is not. In this example $f(z) = 1/(1-z)$ is readily seen to be analytic everywhere except at $z = 1$, which is a "singular point" of f. Theorem 24.2.7 tells us that R is at least 1, and Theorem 24.2.8 tells us that R is no greater than that, because the sum function f is necessarily analytic in $|z| < R$ whereas f is singular at $z = 1$.

In physical terms we can visualize the disk of convergence spreading from the point of expansion, like ripples on a pond, until it encounters a singular point; it can spread no further.

COMMENT 1. Whereas $1/(1-z)$ is meaningful everywhere except at the point $z = 1$, the series $1 + z + z^2 + \cdots$ is meaningful only inside the disk $|z| < 1$ and is therefore only a *partial representation* of the original function. It is interesting to imagine being handed

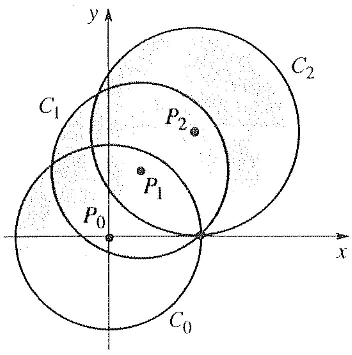


Figure 4. Analytic continuation.

the series $1 + z + z^2 + \cdots$ and not knowing that the sum function is $1/(1 - z)$. Applying the ratio test to the series, we ascertain that it converges only in $|z| < 1$. Nevertheless, if we wish to compute values of the mysterious function (of which the series is a partial representation) *outside* of $|z| < 1$, we can (at least in principle) proceed as follows. Denote the function, of which $1 + z + z^2 + \cdots$ is a partial representation, as $f(z)$. Then we can compute $f, f', f'', \ldots$ at some point P_1, of our choosing, within $|z| < 1$ (Fig. 4). With these values in hand, we can expand f in Taylor series about P_1, obtaining

$$f(z) = f(P_1) + f'(P_1)(z - P_1) + \frac{f''(P_1)}{2!}(z - P_1)^2 + \cdots. \tag{29}$$

Then, applying the ratio test (or some other convergence test) to this new Taylor series, we would find that it converges inside the circle C_1. [Since we are pretending not to know that f happens to be $1/(1-z)$, we do not know about the singular point at $z = 1$.] Similarly, we can use the representation (29) to compute $f, f', f'', \ldots$ at some other point P_2 (Fig. 4) and thus derive still another Taylor series representation of f, this time valid inside C_2, and so on. This "stepping-stone" process is an example of **analytic continuation.*** Incidentally, the best we could possibly do in this example, given the series $1+z+\cdots$, is to obtain $f(z) = 1/(1-z)$, which is the most complete possible analytic continuation of this particular series.

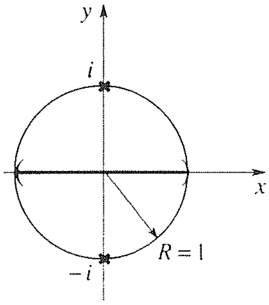


Figure 5. Disk of convergence.

COMMENT 2. Consider, instead, the function

$$g(z) = \frac{1}{1 + z^2}. \tag{30}$$

We can see that g is analytic everywhere except at $z = \pm i$. Thus, we know in advance, even before generating its Taylor series about $z = 0$, that the series will converge in $|z| < 1$ and diverge in $|z| \geq 1$ (Fig. 5). To generate the Taylor series we can, of course, use (15). However, it is easier to recall the geometric series

$$\frac{1}{1 - t} = 1 + t + t^2 + \cdots \tag{31}$$

and to notice that (30) is of that form, with $t = -z^2$. Then, without further ado, (31) gives

$$\frac{1}{1 + z^2} = 1 - z^2 + z^4 - z^6 + \cdots. \tag{32}$$

Since (31) converges in $|t| < 1$, and t is related to z by $t = -z^2$, it follows from $|z| = \sqrt{|t|}$ that (32) converges in $|z| < 1$, as we predicted from the fact that g has singularities at $\pm i$.

If we restrict z to lie on the real axis, then (32) becomes the real Taylor series

$$\frac{1}{1 + x^2} = 1 - x^2 + x^4 - x^6 + \cdots \qquad (|x| < 1). \tag{33}$$

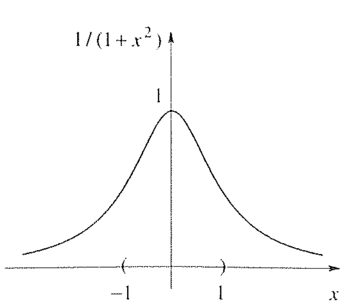


Figure 6. Graph of $1/(1 + x^2)$.

If one encounters (33) in studying the real variable calculus, the limitation $|x| < 1$ must surely seem paradoxical. For why should the series diverge outside $|x| < 1$ if the function $1/(1 + x^2)$ is so beautifully behaved (e.g., infinitely differentiable) for *all* x (Fig. 6)?

*For discussion of analytic continuation, see for example E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis* (Englewood Cliffs, NJ: Prentice Hall, 1976), or Konrad Knopp, *Theory of Functions*, Part I (New York: Dover, 1945).

However, our study of complex function theory now renders the situation transparent: the restriction $|x| < 1$ in (33) occurs because the function $1/(1+z^2)$ has singularities *off the real axis* at $z = +i$ and $z = -i$. ▮

EXAMPLE 7. As a second example in which the results of real variable theory appear paradoxical, consider the function

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{34}$$

the graph of which is shown in Fig. 7. Besides $f(0) = 0$, one can show (which is left for the exercises) that $f'(0) = f''(0) = f'''(0) = \cdots = 0$ so the Taylor series of f about $x = 0$ is simply $0 + 0 + 0 + \cdots$. Surely, the latter converges for *all* x (namely, to zero), but it does not *represent* the given f anywhere except at $x = 0$, the only point at which $f(x)$ is exactly zero. Why this striking failure of the Taylor series? After all, f is infinitely differentiable for all x and hence seems quite well behaved.

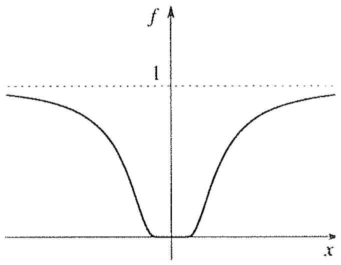


Figure 7. Graph of f given by (3

Again the "paradox" is resolved by examining the behavior of

$$f(z) = \begin{cases} e^{-1/z^2}, & z \neq 0 \\ 0, & z = 0 \end{cases} \tag{35}$$

in the complex plane. Specifically, observe that if z is on the real axis, but $z \neq 0$, then $f(z) = e^{-1/x^2} \to 0$ as $x \to 0$; but if $z \to 0$ along the *imaginary axis*, then $f(z) = e^{-1/(iy)^2} = e^{1/y^2} \to \infty$ as $y \to 0$. Thus, $f(z)$ is not even *continuous* at $z = 0$, let alone analytic. Hence, its Taylor series expansion about $z = 0$ is doomed to converge to f only at the expansion point $z = 0$. ▮

EXAMPLE 8. *Undetermined Coefficients.* In this final example we emphasize the practical matter of how to generate the Taylor series of a given function in a convenient manner. To illustrate, consider the expansion of

$$f(z) = \frac{e^z}{\cos z} \tag{36}$$

about $z = 0$. First, observe that f is singular only at the zeros of $\cos z$, namely, at $z = \pm\pi/2, \pm 3\pi/2, \ldots$. Of these, the closest to the origin are $\pm\pi/2$, and these will limit the radius of convergence of the Maclaurin expansion to $R = \pi/2$ (Fig. 8).

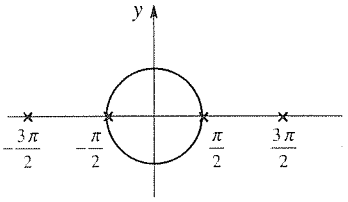


Figure 8. Disk of convergence.

Naturally, we can work out $f(0), f'(0), \ldots$ and use the Taylor series formula. However, the repeated differentiations grow increasingly tedious and we wonder whether we might, instead, benefit from the fact that the expansions of e^z and $\cos z$ are known (i.e., remembered or easily derived). Thus, let us proceed as follows. First, express

$$\frac{e^z}{\cos z} = a_0 + a_1 z + a_2 z^2 + \cdots, \tag{37}$$

where the a_j's are to be determined. Multiplying both sides of (37) by $\cos z$ and then recalling the expansions of e^z and $\cos z$, write

$$1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots = (a_0 + a_1 z + a_2 z^2 + \cdots)\left(1 - \frac{z^2}{2} + \frac{z^4}{24} + \cdots\right). \tag{38}$$

The first series on the right converges in $|z| < \pi/2$, and the second (the cosine series) converges in $|z| < \infty$ so it is justified (Theorem 24.2.10, below) to multiply them term by term, in $|z| < \pi/2$, and to rearrange the result in ascending powers of z. That step gives

$$1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \cdots = a_0 + a_1 z + \left(a_2 - \frac{1}{2}a_0\right) z^2 + \left(a_3 - \frac{1}{2}a_1\right) z^3 + \cdots. \quad (39)$$

Finally, equating coefficients of like powers of z (Theorem 24.2.9, below), we obtain $a_0 = 1$, $a_1 = 1$, $a_2 = 1$, and $a_3 = \frac{2}{3}$ so

$$\frac{e^z}{\cos z} = 1 + z + z^2 + \frac{2}{3}z^3 + \cdots \quad (40)$$

in $|z| < \pi/2$. ▮

In the preceding example we discovered a need for additional theorems covering the manipulation of power series. For instance, we solved for $a_0, a_1, \ldots$ by equating coefficients of like powers of z on the left- and right-hand sides of (39). Surely

$$\sum_{n=0}^{\infty} a_n(z-a)^n = \sum_{n=0}^{\infty} b_n(z-a)^n \quad (41)$$

holds if $a_n = b_n$ for each n, but is a_n *necessarily* equal to b_n, for each n, for (41) to hold? The answer is yes:

THEOREM 24.2.9 *Uniqueness of Power Series*
If $\sum_{n=0}^{\infty} a_n(z-a)^n$ and $\sum_{n=0}^{\infty} b_n(z-a)^n$ both converge in $|z| < R$, then (41) holds in $|z| < R$ if and only if $a_n = b_n$ for each $n = 0, 1, 2, \ldots$.

Proof: Since the sum functions of the two series are identical, say $f(z)$, it follows from Theorem 24.2.8 that $a_n = f^{(n)}(a)/n!$ and $b_n = f^{(n)}(a)/n!$. Hence, $a_n = b_n$ for each n. ■

Also, in proceeding from (38) to (39) we multiplied two power series term by term and rearranged the resulting terms in ascending powers. The following theorem justifies such manipulation.

THEOREM 24.2.10 *Termwise Product of Power Series*
If $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges to $f(z)$ in $|z-a| < R_1$ and $\sum_{n=0}^{\infty} b_n(z-a)^n$ converges to $g(z)$ in $|z-a| < R_2$, then the termwise product

$$\left(\sum_{n=0}^{\infty} a_n(z-a)^n\right)\left(\sum_{n=0}^{\infty} b_n(z-a)^n\right)$$

$$= a_0b_0 + (a_0b_1 + a_1b_0)(z-a) + (a_0b_2 + a_1b_1 + a_2b_0)(z-a)^2 + \cdots$$
$$= \sum_{n=0}^{\infty}(a_0b_n + a_1b_{n-1} + \cdots + a_nb_0)(z-a)^n \tag{42}$$

converges to $f(z)g(z)$ in $|z-a| < \min(R_1, R_2)$.

The resulting series, on the right-hand side of (42), is known as the **Cauchy product** of the two original series.

EXAMPLE 9. Obtain the Cauchy product of the Maclaurin expansions of $1/(1-z)$ and $1/(1+2z)$.

$$\begin{aligned}\frac{1}{1-z}\,\frac{1}{1+2z} &= (1+z+z^2+\cdots)(1-2z+4z^2-\cdots)\\ &= (1)(1) + [(1)(-2)+(1)(1)]z + [(1)(4)+(1)(-2)+(1)(1)]z^2+\cdots\\ &= 1 - z + 3z^2 + \cdots,\end{aligned} \tag{43}$$

which result does agree with the Maclaurin expansion of the product function $1/(1+z-2z^2)$. Since the original series converge in $|z|<1$ and $|z|<1/2$, respectively, their Cauchy product converges to $1/(1+z-2z^2)$ in $|z|<1/2$. ▮

We give three more theorems about the manipulation of power series. Incidentally, we now know that when we speak of convergent power series we might as well call them Taylor series since Theorem 24.2.8 revealed that a convergent power series *is* a Taylor series, the Taylor series of the sum function.

THEOREM 24.2.11 *Termwise Linear Combination of Power Series*
If $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges to $f(z)$ in $|z-a| < R_1$ and $\sum_{n=0}^{\infty} b_n(z-a)^n$ converges to $g(z)$ in $|z-a| < R_2$, then

$$\alpha\sum_{n=0}^{\infty} a_n(z-a)^n + \beta\sum_{n=0}^{\infty} b_n(z-a)^n = \sum_{n=0}^{\infty}(\alpha a_n + \beta b_n)(z-a)^n \tag{44}$$

converges to $\alpha f(z) + \beta g(z)$ in $|z-a| < \min(R_1, R_2)$.

Proof:

$$\lim_{N\to\infty}\sum_{n=0}^{N}(\alpha a_n + \beta b_n)(z-a)^n = \alpha\lim_{N\to\infty}\sum_{n=0}^{N} a_n(z-a)^n + \beta\lim_{N\to\infty}\sum_{n=0}^{N} b_n(z-a)^n$$
$$= \alpha f(z) + \beta g(z),$$

where each of the series on the right-hand side converge. ■

In particular, the cases $\alpha = 1$ and $\beta = \pm 1$ show that the power series can be added and subtracted termwise within the smaller disk of convergence.

THEOREM 24.2.12 *Termwise Differentiation and Integration of Power Series*
If $\sum_{n=0}^{\infty} a_n(z-a)^n$ converges to $f(z)$ in $|z-a| < R$, then

(a) the series $\sum_{n=1}^{\infty} na_n(z-a)^{n-1}$, obtained by termwise differentiation, converges to $f'(z)$ in $|z-a| < R$, and

(b) the series $\sum_{n=0}^{\infty} \dfrac{a_n}{n+1}(z-a)^{n+1}$, obtained by termwise integration, converges in $|z-a| < R$.

Proof of (a): Since the given series converges to $f(z)$, it must be the Taylor series of f (Theorem 24.2.8); hence, $a_n = f^{(n)}(a)/n!$. We learned, from our discussion of the Cauchy integral formula, that if f is analytic in a given region then so is f', f'', f''', and so on. Thus, $f'(z)$ can be represented by a Taylor series in $|z-a| < R$:

$$\begin{aligned} f'(z) &= \sum_{n=0}^{\infty} \frac{f'^{(n)}(a)}{n!}(z-a)^n = \sum_{n=0}^{\infty}(n+1)\frac{f^{(n+1)}(a)}{(n+1)!}(z-a)^n \\ &= \sum_{n=0}^{\infty}(n+1)a_{n+1}(z-a)^n = \sum_{n=1}^{\infty} na_n(z-a)^{n-1}, \end{aligned}$$

as was to be shown. ■

Closure. In Section 24.2.1 we introduce the concept of complex series, and cover several tests for the convergence of complex series of constants and complex power series. Whereas a real power series $\sum a_n(x-a)^n$ converges in an interval $|x-a| < R$ on the x axis, a complex power series $\sum a_n(z-a)^n$ converges in a *disk* $|z-a| < R$ in the z plane.

In Section 24.2.2 we focus on the Taylor series of a function of a complex variable and use the Cauchy integral formula to derive the Taylor series of $f(z)$ about $z = a$ and to show that the series converges to $f(z)$, and hence "represents f," in the largest possible disk, centered at a, throughout which f is analytic. Since the theory is really an extension of the real variable theory, the expansions of the various elementary functions carry over intact, with x's changed to z's and intervals of convergence changed to disks of convergence. For instance,

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n \qquad \text{in } |z| < 1, \tag{45}$$

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} \qquad \text{in } |z| < \infty, \qquad (46)$$

$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} \qquad \text{in } |z| < \infty, \qquad (47)$$

$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} \qquad \text{in } |z| < \infty, \qquad (48)$$

$$\sinh z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} \qquad \text{in } |z| < \infty, \qquad (49)$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} \qquad \text{in } |z| < \infty. \qquad (50)$$

With Theorem 24.2.8 the distinction between power series and Taylor series virtually disappears since a convergent power series is indeed the Taylor series of its sum function.

Examples 6 and 7 reveal that the Taylor series of real variable theory affords only a partial view of the whole, and can therefore sometimes appear paradoxical. For instance, the Taylor series of $1/(1+x^2)$ about $x = 0$ converges only in $-1 < x < 1$, even though the function is well behaved at $x = \pm 1$ and, indeed, is infinitely differentiable for all x. The source of the paradox is revealed by considering instead the Taylor series of $1/(1+z^2)$ about $z = 0$ in the complex plane; namely, it is the presence of singularities at $z = \pm i$, *off* of the real axis, that limits the disk of convergence in the z plane and hence the interval of convergence on the x axis.

Example 8, on a method of undetermined coefficients, reveals a general need for more theorems on the manipulation of power series, and we close this section with several such theorems, the gist of which is that convergent power series can be added, subtracted, multiplied, differentiated, and integrated termwise within their common disk of convergence.

EXERCISES 24.2

1. In the second paragraph of Section 24.2.1 we state that $\sum c_n$ converges if and only if $\sum a_n$ and $\sum b_n$ do, where $c_n = a_n + ib_n$. Prove that claim, using the "ϵ definition" of convergence given in the first paragraph.

2. Cite the justification for the first inequality in (2).

3. Prove Theorem 24.2.4 for the case where $L > 1$.

4. Use any of Theorems 24.2.1–4 to prove Theorem 24.2.5.

5. Use any of the theorems of Section 24.2.1 to determine the convergence or divergence of the following complex series.

(a) $\displaystyle\sum_{n=1}^{\infty} \frac{n}{(2+i)^n}$ (b) $\displaystyle\sum_{n=5}^{\infty} \frac{n^{50}}{3^n}$

(c) $\displaystyle\sum_{n=2}^{\infty} \frac{1}{(2i)^n \ln n}$ (d) $\displaystyle\sum_{n=0}^{\infty} \left(\frac{1+n}{2+n}\right)^3$

(e) $\displaystyle\sum_{n=1}^{\infty} \frac{(1+3i)^n}{n^{100}}$ (f) $\displaystyle\sum_{n=1}^{\infty} n^4 e^{-(5-i)n}$

(g) $\displaystyle\sum_{n=1}^{\infty} e^{-in}$ (h) $\displaystyle\sum_{n=1}^{\infty} (\sin n) \left(\frac{1+i}{2-i}\right)^n$

6. Use any of the theorems of Section 24.2.1 to determine, insofar as possible, the regions of convergence and divergence of the following power series.

(a) $\sum_{n=0}^{\infty} z^{2n}$ (b) $\sum_{n=2}^{\infty} n^2(z-3)^n$

(c) $\sum_{n=0}^{\infty} n!(z+5)^n$ (d) $\sum_{n=4}^{\infty} e^n(z+i)^n$

(e) $\sum_{n=1}^{\infty} e^{-n}z^n$ (f) $\sum_{n=0}^{\infty} (n^{100}/n!)z^n$

(g) $\sum_{n=0}^{\infty} e^{in}z^n$ (h) $\sum_{n=0}^{\infty} \frac{\cos n}{n^2+1}z^n$

(i) $\sum_{n=1}^{\infty} [(2-i)z]^n$ (j) $\sum_{n=3}^{\infty} (e^{n^2}/n!)z^n$

7. For the function $f(x)$ defined by (34), it is claimed that derivatives of all orders exist and are zero at $x=0$. Verify that claim for the derivatives of first and second order.

8. Is the following power series the Taylor series of some function? If so, in what region does the series represent the function; if not, why not?

(a) $\sum_{n=1}^{\infty} (-1)^{n+1}n(z-1)^n$ (b) $\sum_{n=0}^{\infty} \frac{z^{n-1}}{4^{n+1}}$

(c) $\sum_{n=4}^{\infty} \left(\frac{z+i}{1+i}\right)^{2n}$ (d) $1+z^3$

9. What is the radius of convergence of the Taylor series of $1/(z^2-3z+2)$ about

(a) $z=0$ (b) $z=3i$ (c) $z=1-5i$ (d) $z=5-i$

10. What is the radius of convergence of the Taylor series of $(z^2-3z+2)/(z^2-2z+3i+1)$ about

(a) $z=0$ (b) $z=10i$ (c) $z=2-5i$ (d) $z=20$

11. Obtain the Taylor series of the given function, about $z=a$, and give its radius of convergence R.

(a) $\sin z$, $a=0$ (b) $\sin z$, $a=2-i$

(c) $\cos 2z$, $a=3i$ (d) e^{z^6}, $a=0$

(e) $\frac{1}{i+z}$, $a=0$ (f) $\frac{z^3}{2-iz}$, $a=0$

(g) $\sin z^8$, $a=0$ (h) z^3, $a=-2i$

(i) $\frac{1}{1+2z^{35}}$, $a=0$ (j) z^2-iz, $a=2i$

12. (*Binomial series*) (a) Derive the **binomial series**

$$\begin{aligned}\frac{1}{(1-z)^m} &= 1+mz+\frac{m(m+1)}{2!}z^2 \\ &\quad +\frac{m(m+1)(m+2)}{3!}z^3+\cdots \\ &= \sum_{n=0}^{\infty}\frac{(m+n-1)!}{(m-1)!\,n!}z^n \qquad (|z|<1)\end{aligned} \tag{12.1}$$

for any positive integer m.
(b) Use (12.1) to obtain the Taylor series of $1/(3-z)^2$ about $z=i$. Identify its disk of convergence. HINT: Write

$$\frac{1}{(3-z)^2}=\frac{1}{[3-(z-i)-i]^2}=\frac{1}{(3-i)^2\left[1-\left(\frac{z-i}{3-i}\right)\right]^2} \tag{12.2}$$

and use $(z-i)/(3-i)$ in place of z in (12.1).

13. Use (12.1) in Exercise 12 to obtain the Taylor expansions

(a) $\frac{1}{(2z+1)^3}=\frac{1}{2}\sum_{n=0}^{\infty}(-1)^n(n+1)(n+2)2^nz^n$
in $|z|<\frac{1}{2}$

(b) $\frac{1}{(2z+1)^3}=\frac{1}{250}\sum_{n=0}^{\infty}(-1)^n(n+1)(n+2)\left(\frac{2}{5}\right)^n(z-2)^n$
in $|z-2|<\frac{5}{2}$

(c) $\frac{1}{z^2-z-6}=\frac{1}{20}\sum_{n=0}^{\infty}\frac{(-4)^{n+1}-1}{4^n}(z+1)^n$
in $|z+1|<1$

14. Let $\sqrt{z}$ be defined by a principal-value branch cut (i.e., with the origin and negative x axis deleted and $-\pi<\theta<\pi$). Work out the first several terms of the Taylor series, about the given point, and give the region of validity of the Taylor series representation–that is, the region in which the series converges to the given function.

(a) $a=1$ (b) $a=-i$ (c) $a=i$
(d) $a=-1-i$ (e) $a=-2+i$ (f) $a=2+i$

15. Determine the coefficients of the next two terms in (40) (i.e., the z^4 and z^5 terms).

16. Use the method of undetermined coefficients to determine the first several terms of the Maclaurin series of the given function, and give its region of convergence.

(a) $\tan z = \dfrac{\sin z}{\cos z}$ (b) $\sec z = \dfrac{1}{\cos z}$

(c) $\operatorname{cosec} z = \dfrac{1}{\sin z}$ (d) $\dfrac{1+z}{1+2z+3z^2}$

(e) $\dfrac{3-z}{2+3z^2+z^4}$ (f) $\dfrac{e^z}{\sin 2z}$

(g) $\dfrac{1}{2-\sin z}$ (h) $\dfrac{1}{3+\cos z}$

17. Show that $f'(0) = f''(0) = \cdots = 0$, as claimed in Example 7. HINT: Use the difference quotient definition of the derivative.

24.3 Laurent Series

Consider the function

$$f(z) = \frac{1}{(z-1)(z-2i)}, \tag{1}$$

which is analytic for all z except at the two (singular) points $z = 1$ and $z = 2i$. If we develop a Taylor expansion about $z = 0$, that expansion will be valid [that is, will converge to $f(z)$] in the disk $|z| < 1$. Thus, expanding about $z = 0$, the region $|z| \geq 1$ is inaccessible to us. However, there exists a more general representation known as a *Laurent series*, which includes the Taylor series as a special case and which permits expansions in any *annulus* throughout which f is analytic. Thus, for the f given above, there would be three possible Laurent expansions about $z = 0$: one in $2 < |z| < \infty$, one in $1 < |z| < 2$, and one in $|z| < 1$ (Fig. 1a). (The latter would simply be the *Taylor series*.) In fact, with the Laurent series one is even able to expand about a singular point! For the function f given above, for example, we can expand about $z = 1$ in $0 < |z-1| < \sqrt{5}$ or in $\sqrt{5} < |z-1| < \infty$ (Fig. 1b); similarly, we could expand about the other singular point $z = 2i$.

The relevant theorem is as follows.

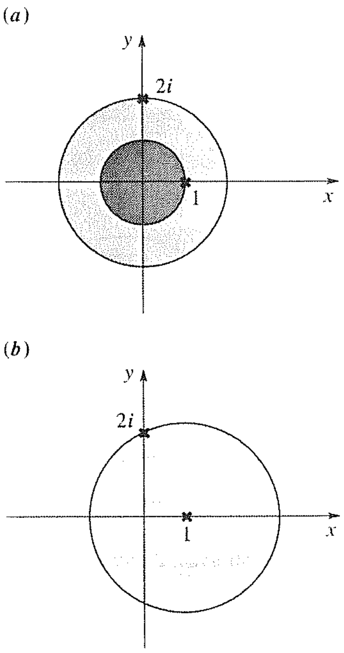


Figure 1. Laurent series regions for f given by (1): (a) expansions about $z = 0$, (b) expansions about $z = 1$.

THEOREM 24.3.1 *Laurent Series*
Let D be the closed region between, and including, concentric circles C_i, C_o, with their centers at $z = a$. If $f(z)$ is analytic in D, then it admits the **Laurent series** representation

$$\boxed{f(z) = \sum_{n=-\infty}^{\infty} c_n (z-a)^n} \tag{2}$$

in D, with the c_n's given uniquely by

$$c_n = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - a)^{n+1}}\, d\zeta, \tag{3}$$

where C is any piecewise smooth simple closed counterclockwise path in D.

Partial Proof: The circles C_i and C_o are shown in Fig. 2a. Our derivation of (2) will be similar to our derivation of the Taylor series formula (15) in Section 24.2, which derivation we urge you to review before proceeding. Once again our starting point is the Cauchy integral formula, but since the region of interest (the shaded annulus in Fig. 2a) is not bounded by a single piecewise smooth simple closed curve we cannot yet apply that formula. Thus, we first introduce a slit through the annulus, so as to create the piecewise smooth simple curve $C_1 + C_2 + C_3 + C_4$ shown in Fig. 2b, where C_1 is the circle C_o taken counterclockwise, C_3 is the circle C_i taken clockwise, and C_2, C_4 are abutting and oppositely oriented, where C_2 and C_4 are shown as slightly separated for clarity. Further, $f(z)$ is analytic inside and on that contour so Cauchy's integral formula applies and gives

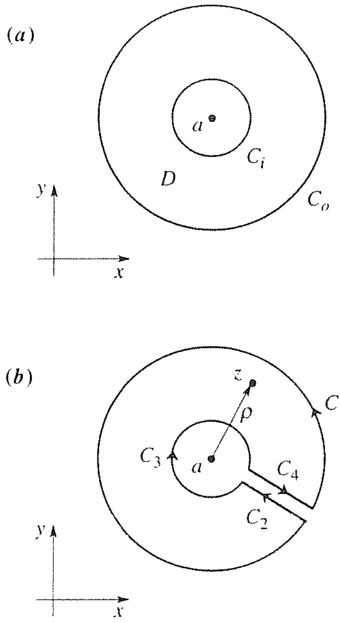


Figure 2. Deriving (2).

$$\begin{aligned} f(z) &= \frac{1}{2\pi i}\oint_{C_1+C_2+C_3+C_4} \frac{f(\zeta)}{\zeta - z}\,d\zeta \\ &= \frac{1}{2\pi i}\oint_{C_1}\frac{f(\zeta)}{\zeta - z}\,d\zeta + \frac{1}{2\pi i}\oint_{C_3}\frac{f(\zeta)}{\zeta - z}\,d\zeta, \end{aligned} \tag{4}$$

where the second equality in (4) holds because the contributions from C_2 and C_4 are equal and opposite and hence cancel.

Considering the C_1 integral first, express

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a}\,\frac{1}{1 - \dfrac{z-a}{\zeta - a}} = \frac{1}{\zeta - a}\sum_{n=0}^{\infty}\left(\frac{z-a}{\zeta-a}\right)^n, \tag{5}$$

where the first step in (5) is arranged so that the resulting series, on the right-hand side, converges for all ζ on C_1 because $\left|\frac{z-a}{\zeta-a}\right| = |z-a|\,/\,|\zeta - a| < 1$ for all ζ on C_1. Thus,

$$\begin{aligned} \frac{1}{2\pi i}\oint_{C_1}\frac{f(\zeta)}{\zeta - z}\,d\zeta &= \frac{1}{2\pi i}\oint_{C_1}\frac{f(\zeta)}{\zeta - a}\sum_{n=0}^{\infty}\left(\frac{z-a}{\zeta-a}\right)^n d\zeta \\ &= \sum_{n=0}^{\infty}\left[\frac{1}{2\pi i}\oint_{C_1}\frac{f(\zeta)}{(\zeta-a)^{n+1}}\,d\zeta\right](z-a)^n \\ &= \sum_{n=0}^{\infty} c_n (z-a)^n, \end{aligned} \tag{6}$$

where

$$c_n = \frac{1}{2\pi i}\oint_{C_1}\frac{f(\zeta)}{(\zeta - a)^{n+1}}\,d\zeta.$$

The path C_1 can be deformed to any piecewise smooth simple closed counterclockwise path C lying entirely in D because the integrand is analytic between and on C_1 and C so let us re-express c_n as

$$c_n = \frac{1}{2\pi i}\oint_{C}\frac{f(\zeta)}{(\zeta - a)^{n+1}}\,d\zeta. \tag{7}$$

Next, consider the C_3 integral in (4). This time, express

$$\frac{1}{\zeta - z} = -\frac{1}{z-a}\,\frac{1}{1-\dfrac{\zeta - a}{z-a}} = -\frac{1}{z-a}\sum_{n=0}^{\infty}\left(\frac{\zeta - a}{z-a}\right)^n, \tag{8}$$

where the series, on the right-hand side converges for all ζ on C_3 because $\left|\frac{\zeta-a}{z-a}\right| = |\zeta - a| \,/\, |z-a| < 1$ for all ζ on C_3. Thus,

$$\begin{aligned}\frac{1}{2\pi i}\oint_{C_3}\frac{f(\zeta)}{\zeta - z}\,d\zeta &= -\frac{1}{2\pi i}\oint_{C_3}\frac{f(\zeta)}{z-a}\sum_{n=0}^{\infty}\left(\frac{\zeta-a}{z-a}\right)^n d\zeta\\ &= \sum_{n=0}^{\infty}\left[-\frac{1}{2\pi i}\oint_{C_3} f(\zeta)(\zeta-a)^n\,d\zeta\right](z-a)^{-(n+1)}\\ &= \sum_{n=-1}^{-\infty}\left[\frac{1}{2\pi i}\oint_{-C_3}\frac{f(\zeta)}{(\zeta-a)^{n+1}}\,d\zeta\right](z-a)^n\\ &= \sum_{n=-1}^{-\infty}\left[\frac{1}{2\pi i}\oint_{C}\frac{f(\zeta)}{(\zeta-a)^{n+1}}\,d\zeta\right](z-a)^n\\ &= \sum_{n=-1}^{-\infty} c_n(z-a)^n. \end{aligned}\tag{9}$$

In the third equality in (9) we shifted the summation index so as to obtain powers of $z-a$, as are present in (6). In the fourth equality we deformed $-C_3$ (i.e., the contour C_3 with its direction reversed so as to be counterclockwise) to the same counterclockwise contour C as in (7). Observe that the c_n's in (9) are defined by the same integral expression as those in (7). Thus, putting (6), (7), and (9) into (4) does give the series (2), with the c_n's given by (3).

Observe that we have not justified the interchanges in the order of the infinite summation and the integration, which occurred in the second equality in (6) and in the second equality in (9). We could do that in either of two ways. First, instead of using infinite series, in (5) and (8), we could have used finite series with remainder terms [as we did in equations (18)–(20) in Section 24.2], and then shown, by ML bounds, that the remainder term integrals tend to zero as $n \to \infty$. We have omitted those steps for the sake of brevity. Second, we could have justified the limit interchanges by showing that the infinite series in (5) and (8) converge *uniformly*, but–again for brevity–we have not included the discussion of *uniform convergence* that would be needed. ■

Observe that the sum in (2) is unlike the power series and Taylor series of Section 24.2 in that it contains both positive and negative powers: n runs from $-\infty$ to $+\infty$. Since such a sum is new, for us, we need to define it, just as we defined

$\sum_{n=1}^{\infty} c_n$ as $\lim_{N\to\infty}\sum_{n=1}^{N} c_n$ in Section 24.2.1. Reviewing the foregoing proof, we see that (2) came into being as the sum of two "ordinary" series,

$$\sum_{n=-\infty}^{\infty} c_n(z-a)^n = \sum_{n=0}^{\infty} c_n(z-a)^n + \sum_{n=-1}^{-\infty} c_n(z-a)^n \tag{10}$$

so it follows that we are to understand the Laurent series in (2) as

$$\sum_{n=-\infty}^{\infty} c_n(z-a)^n \equiv \lim_{M\to\infty}\sum_{n=0}^{M} c_n(z-a)^n + \lim_{N\to\infty}\sum_{n=1}^{N} c_{-n}(z-a)^{-n}, \tag{11}$$

where M and N tend to infinity independently.

Observe further that $f(z)$ has been assumed analytic only in the annulus D. Thus, Theorem 24.3.1 allows for singularities inside C_i and outside C_o.

If, in fact, there are no singular points inside C_i, then (3) gives

$$c_n = \begin{cases} \dfrac{f^{(n)}(a)}{n!}, & n = 0, 1, 2, \dots \\ 0, & n = -1, -2, \dots \end{cases} \tag{12}$$

by the generalized Cauchy integral formula and Cauchy's theorem, respectively, and the Laurent series (2) reduces to the Taylor series

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(z-a)^n, \tag{13}$$

as indeed it should.

However, if $f(z)$ *does* have singular points inside C_i (so that one does not merely have a Taylor series), then (12) does not hold, and we are "stuck" with the integral expression of the c_n's that is given in (3). Thus, whereas Taylor series coefficients are the easily computed quantities $f^{(n)}(a)/n!$, Laurent series coefficients are given by the unwieldy integral expression (3). Do we really need to evaluate that integral to obtain the c_n's? No, *practically speaking we are always able to avoid (3) in developing Laurent series, as illustrated in the examples to follow.*

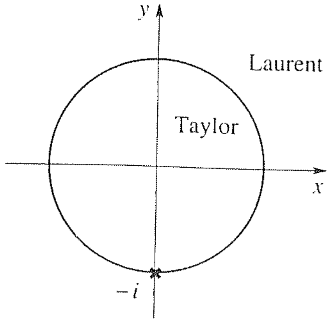


Figure 3. Two possible expansions about $z = 0$.

EXAMPLE 1. Obtain all possible expansions of

$$f(z) = \frac{1}{z+i} \tag{14}$$

about $z = 0$. Since f is singular only at $z = -i$, it admits exactly two possible expansions about $z = 0$, a Taylor series in $|z| < 1$ and a Laurent series in the annulus $1 < |z| < \infty$ (Fig. 3). The Taylor series is

$$\begin{aligned} \frac{1}{z+i} &= \frac{1}{i}\frac{1}{1+\frac{z}{i}} = -i\frac{1}{1-iz} \\ &= -i[1 + (iz) + (iz)^2 + \cdots] \\ &= -i + z + iz^2 - \cdots, \qquad (|z| < 1) \end{aligned} \tag{15}$$

where we factored out the i in the denominator simply so that we could use the remembered geometric series formula $1/(1-t) = 1+t+t^2+\cdots$ $(|t| < 1)$. Alternatively, we could have applied the Taylor series formula $f(z) = f(a) + f'(a)(z-a) + \cdots$ to $f(z) = 1/(z+i)$.

Next, determine the Laurent expansion in $1 < |z| < \infty$. This time, write

$$\frac{1}{z+i} = \frac{1}{z}\frac{1}{1+\dfrac{i}{z}}. \tag{16}$$

Since we are expanding about $z = 0$, the desired Laurent series will proceed in powers of z. The first factor in (16), $1/z$, is already in powers of z, so leave it intact. In the second factor, set $t = i/z$ and observe that $|t| = |i/z| = 1/|z| < 1$ because $|z| > 1$, so we can write the *Taylor* expansion

$$\frac{1}{1+i/z} = \frac{1}{1+t} = 1 - t + t^2 - \cdots \qquad (|t| < 1) \tag{17}$$

in the variable t. With the expansion accomplished, we revert to the z variable by putting $t = i/z$. Thus,

$$\begin{aligned}\frac{1}{1+i/z} &= 1 - \frac{i}{z} + \left(\frac{i}{z}\right)^2 - \cdots \\ &= 1 - \frac{i}{z} - \frac{1}{z^2} + \cdots, \qquad (1 < |z| < \infty)\end{aligned} \tag{18}$$

so (16) becomes

$$f(z) = \frac{1}{z+i} = \left(\frac{1}{z}\right)\left(1 - \frac{i}{z} - \frac{1}{z^2} + \cdots\right). \tag{19}$$

Since the first factor is valid in the annulus $0 < |z| < \infty$ (i.e., the plane with the origin removed because $1/z$ is undefined at $z = 0$), and the second is valid in the annulus $1 < |z| < \infty$, the product

$$f(z) = \frac{1}{z} - \frac{i}{z^2} - \frac{1}{z^3} + \cdots \tag{20}$$

is valid in the overlap annulus $1 < |z| < \infty$ and is the desired Laurent expansion there. More generally, Laurent series contain both positive and negative powers; (20) happens to contain only negative powers.

COMMENT 1. The expansions (15) and (20) are about the same point, $z = 0$, yet they are different from each other. That difference does not violate uniqueness for they are valid in different annuli: (15) is the unique expansion in $|z| < 1$, and (20) is the unique expansion in $1 < |z| < \infty$.

COMMENT 2. Suppose we wish to expand about a point other than $z = 0$, say $z = 2$. We can obtain a Taylor series in $|z-2| < \sqrt{5}$, or a Laurent series in $\sqrt{5} < |z-2| < \infty$ (Fig. 4). Let us derive the latter. It is convenient, though not essential, to shift the origin to the point of expansion by settting $t = z - 2$; thus, t is the "vector" from 2 to any point z in the annulus (Fig. 4). Then

$$\frac{1}{z+i} = \frac{1}{t+2+i} = \frac{1}{t}\frac{1}{1+\dfrac{2+i}{t}}$$

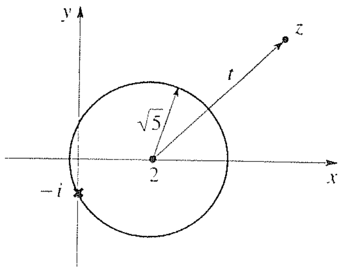


Figure 4. Expanding about $z = 2$.

$$= \frac{1}{t}\left[1 - \frac{2+i}{t} + \frac{(2+i)^2}{t^2} - \cdots\right]$$
$$= \frac{1}{z-2} - (2+i)\frac{1}{(z-2)^2} + (3+4i)\frac{1}{(z-2)^3} - \cdots \tag{21}$$

is the desired expansion, valid in $\sqrt{5} < |z| < \infty$.

COMMENT 3. As noted above, we can even expand about a singular point. In the present example, the Laurent expansion about the singular point $-i$ is simply

$$f(z) = \frac{1}{z+i}; \tag{22}$$

that is, $1/(z+i)$ is already a Laurent expansion about $-i$ because it proceeds in powers of $z+i$, a one-term expansion, just as $5x^3$ is a one-term Taylor expansion about $x = 0$. ▮

Again, we did *not* use (3) to determine the c_n coefficients in the Laurent series (20). Rather, the rearrangement accomplished in (17) enabled us to recast the problem in terms of a *Taylor* series expansion in the new variable t. That idea, recasting the problem in terms of Taylor series, will be relied on exclusively in our derivations of Laurent series.

EXAMPLE 2. The function

$$f(z) = \frac{1}{\sin z} \tag{23}$$

is singular at the zeros of $\sin z$, namely, at $z = 0, \pm\pi, \pm 2\pi, \ldots$. Derive the Laurent expansion of f about $z = \pi$, in the annulus $0 < |z - \pi| < \pi$. (See Fig. 5, where we have made the outer radius a bit less than π, and the inner radius a bit greater than zero, to emphasize that $0 < |z - \pi| < \pi$.) With $t = z - \pi$, for convenience,

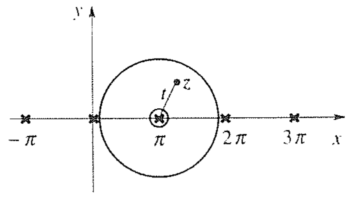


Figure 5. Expansion of $1/\sin z$ in $0 < |z - \pi| < \pi$.

$$\frac{1}{\sin z} = \frac{1}{\sin(t+\pi)} = -\frac{1}{\sin t} = -\frac{1}{t}\frac{t}{\sin t}. \tag{24}$$

To motivate the last equality, observe that $\sin t = t - t^3/3! + \cdots = (t)(1 - t^2/3! + t^4/5! - \cdots)$. The first factor (i.e., t) vanishes at $t = 0$ and causes $1/\sin t$ to be singular there; the second factor is *not* zero at $t = 0$, so $t/\sin t = (1 - t^2/3! + \cdots)^{-1}$ is analytic at $t = 0$. The idea, then, is that the $1/t$ on the right-hand side of (24) is singular at $t = 0$ and is already a one-term Laurent series about $t = 0$. The $t/\sin t$ has been "desingularized" at $t = 0$, by the t in the numerator, but is still singular at $t = \pm\pi, \pm 2\pi, \ldots$ (i.e., at $z = 2\pi, 0, 3\pi, -\pi, \ldots$) so it admits a Taylor series in $0 \le |t| < \pi$, which Taylor series

$$\frac{t}{\sin t} = a_0 + a_1 t + a_2 t^2 + \cdots \tag{25}$$

can be determined, by the method of undetermined coefficients, as

$$\frac{t}{\sin t} = 1 + \frac{1}{6}t^2 + \frac{7}{360}t^4 + \cdots. \tag{26}$$

Thus,

$$\begin{aligned}\frac{1}{\sin z} &= -\left(\frac{1}{t}\right)\left(1+\frac{1}{6}t^2+\frac{7}{360}t^4+\cdots\right)\\ &= -\frac{1}{z-\pi}-\frac{1}{6}(z-\pi)-\frac{7}{360}(z-\pi)^3-\cdots \qquad (27)\end{aligned}$$

is the desired Laurent series. Since the $1/t$ factor converges in the annulus $0 < |t| < \infty$, and the $1 + t^2/6 + \cdots$ factor converges in $0 \le |t| < \pi$, their product converges in the overlap annulus $0 < |t| < \pi$, that is, in $0 < |z - \pi| < \pi$. ▮

EXAMPLE 3. Expand the function

$$f(z) = \frac{1}{z(z-2)}, \qquad (28)$$

which is singular at $z = 0$ and $z = 2$, about $z = i$, in the annulus $\sqrt{5} < |z - i| < \infty$ (Fig. 6). We could expand each of the factors, $1/z$ and $1/(z - 2)$, and multiply their expansions term by term, or we could use partial fractions to express

$$f(z) = -\frac{1}{2}\frac{1}{z}+\frac{1}{2}\frac{1}{z-2}, \qquad (29)$$

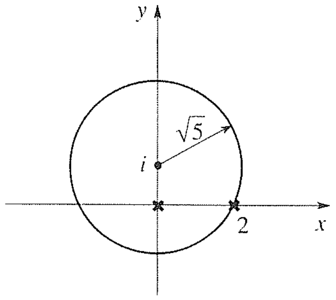


Figure 6. Expansion in $\sqrt{5} < |z - i| < \infty$.

and then add the expansions of $-1/(2z)$ and $1/[2(z - 2)]$. Since addition is easier than multiplication, let us do the latter. With $t = z - i$, we have

$$\begin{aligned}\frac{1}{z} = \frac{1}{t+i} &= \frac{1}{t}\frac{1}{1+\frac{i}{t}} = \frac{1}{t}\left(1-\frac{i}{t}+\frac{i^2}{t^2}-\cdots\right)\\ &= \frac{1}{t}-i\frac{1}{t^2}-\frac{1}{t^3}-\cdots, \qquad (1<|t|<\infty) \qquad (30)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{z-2} = \frac{1}{t+i-2} &= \frac{1}{t}\frac{1}{1+\frac{i-2}{t}}\\ &= \frac{1}{t}\left[1-\frac{i-2}{t}+\frac{(i-2)^2}{t^2}-\cdots\right]\\ &= \frac{1}{t}-(i-2)\frac{1}{t^2}+(3-4i)\frac{1}{t^3}-\cdots, \qquad (\sqrt{5}<|t|<\infty) \qquad (31)\end{aligned}$$

so (29) becomes

$$f(z) = \frac{1}{(z-i)^2}+(2-2i)\frac{1}{(z-i)^3}+\cdots. \qquad (\sqrt{5}<|z-i|<\infty) \qquad (32)$$

COMMENT 1. Working term by term is useful pedagogically because the approach is so concrete, but if possible it is preferable to work with summation notation so as to obtain the full expansion rather than just the first few terms. In the present case we have, from (29)–(31),

$$\begin{aligned} f(z) &= -\frac{1}{2}\frac{1}{t}\sum_{n=0}^{\infty}\left(-\frac{i}{t}\right)^n + \frac{1}{2}\frac{1}{t}\sum_{n=0}^{\infty}\left(-\frac{i-2}{t}\right)^n \\ &= \frac{1}{2}\sum_{n=1}^{\infty}[(2-i)^n - (-i)^n]\frac{1}{(z-i)^{n+1}} \end{aligned} \tag{33}$$

in $\sqrt{5} < |z-i| < \infty$, where we have changed the lower summation limit to 1 simply because $(2-i)^0 - (-i)^0 = 1 - 1 = 0$.

COMMENT 2. The expansion of f about $z = i$ in the other annuli, $0 \le |z-i| < 1$ and $1 < |z-i| < \sqrt{5}$, is left for the exercises. ▮

EXAMPLE 4. Expand

$$f(z) = e^{1/z} \tag{34}$$

about $z = 0$.

Evidently, f is analytic for all $z \neq 0$ and singular for $z = 0$, so the expansion of f about $z = 0$ will be a Laurent expansion valid in the annulus $0 < |z| < \infty$. To obtain that expansion, let $1/z = t$ and observe that $e^{1/z} = e^t$ admits the *Taylor* expansion

$$e^{1/z} = e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots \tag{35}$$

in $|t| < \infty$. Since $|t| < \infty$ corresponds to $|z| > 0$, it follows [by setting $t = 1/z$ in (35)] that the desired Laurent expansion is

$$e^{1/z} = 1 + \frac{1}{z} + \frac{1}{2!}\frac{1}{z^2} + \frac{1}{3!}\frac{1}{z^3} + \cdots \tag{36}$$

in the annulus $0 < |z| < \infty$. ▮

Closure. The focus of this section is that we can expand a function $f(z)$, about any point $z = a$, within any annulus centered at a and throughout which $f(z)$ is analytic. In fact, $f(z)$ need not even be analytic at a, but if it is, then the innermost such annulus is a disk, and the Laurent series in that disk is simpiy the Taylor series of $f(z)$ about a. Besides that theoretical base, provided by Theorem 24.3.1, we emphasize the technique involved in actually generating the Laurent series of a given function in a particular annulus – namely, that the unwieldy integral formula (3) can be avoided, and that the coefficients c_n can be evaluated by recasting the expansion(s) in terms of Taylor series.

EXERCISES 24.3

1. Derive the right-hand side of (26), up to and including the t^6 term, by the method of undetermined coefficients.

2. Expand the function $f(z) = 1/[z(z-2)]$, in Example 3, about $z = i$, in the annulus between the two singular points, and show that

$$f(z) = -\frac{1}{2}\sum_{n=0}^{\infty}(-i)^n(z-i)^{-n-1} - \frac{1}{2}\sum_{n=0}^{\infty}\left(\frac{2+i}{5}\right)^{n+1}(z-i)^n$$

in that annulus.

3. In Example 3 we use partial fractions to re-express (28) as (29), then we expand each of the two terms in (29) and add their series. The result is given in (33). Here, we ask you to work with the product form (28) rather than the sum form (29), and to show that the same final result is obtained. Specifically, show that (28) gives

$$\begin{aligned} f(z) &= \frac{1}{t^2}\sum_{n=0}^{\infty}\left(-\frac{i}{t}\right)^n \sum_{m=0}^{\infty}\left(\frac{2-i}{t}\right)^m \\ &= \frac{1}{t^2}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}(-i)^n(2-i)^m t^{-(m+n)}, \end{aligned} \tag{3.1}$$

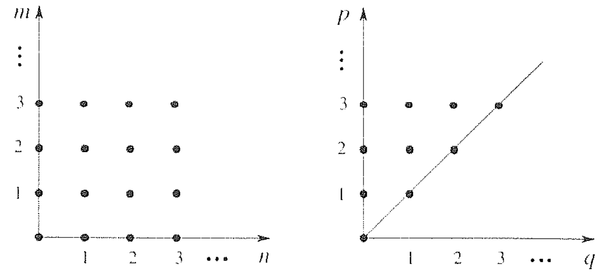
where $t = z - i$. The latter is an iterated sum, and we can handle it in somewhat the same manner as an iterated integral. Since the $m + n$ exponent is "begging" to be a new variable, let us change variables from m, n to p, q according to

$$p = m + n, \qquad q = m \tag{3.2}$$

(or we could use $q = n$; it doesn't matter). Next, show that the regions of summation in the m, n and p, q planes are as shown below, and that

$$f(z) = \sum_{p=0}^{\infty}\left[\sum_{q=0}^{p}(-i)^{p-q}(2-i)^q\right]t^{-(p+2)}. \tag{3.3}$$

Finally, write out the first four terms of (33), and of (3.3), and verify that they agree. In fact, show (33) and (3.3) are identical for *all* terms. HINT: Recall the identity (5.1) in Exercise 5 of Section 4.2. NOTE: It should be clear from this exercise that the partial fraction approach pursued in Example 3 is simpler than the product approach pursued in this exercise.



4. Obtain the first three nonvanishing terms of the Laurent expansion of each of the following.

(a) $\dfrac{1}{z}$ in $1 < |z-i| < \infty$

(b) $\dfrac{1}{z^2+1}$ in $1 < |z| < \infty$

(c) $\dfrac{z^2+3}{z}$ in $0 < |z| < \infty$

(d) $\dfrac{1}{e^z-1}$ in $0 < |z| < 2\pi$

(e) $\dfrac{1}{z(z^3+2)}$ in $0 < |z| < \sqrt[3]{2}$

(f) $\dfrac{1}{z} + \dfrac{z}{z+i}$ in $1 < |z| < \infty$

(g) $\dfrac{1}{z^3}$ in $2 < |z-2| < \infty$

(h) $\dfrac{1}{z^2}$ in $1 < |z+i| < \infty$

(i) $\dfrac{1}{\cos z}$ in $0 < \left|z - \dfrac{\pi}{2}\right| < \pi$

(j) $\tan z$ in $0 < \left|z + \dfrac{\pi}{2}\right| < \pi$

5. Determine all possible Taylor and Laurent expansions about the given point $z = a$, and state their regions of validity.

(a) $\sin\dfrac{1}{z}$, $a = 0$

(b) $\dfrac{1}{z}$, $a = -2$

(c) e^{-1/z^3}, $a = 0$

(d) $\dfrac{z^2+5}{z^2+4}$, $a = -1$

(e) $\dfrac{1+z}{2+z}$, $a = -i$

(f) $\dfrac{\sin z}{z^4}$, $a = 0$

(g) $\dfrac{\cos 2z}{(z+i)^2}$, $a = -i$

(h) e^{-z^2}, $a = 0$

(i) $e^{-1/z}$, $a = 0$ (j) $\dfrac{1}{z(z^2+1)}$, $a = 1$

(k) $\dfrac{1}{z^2+iz+2}$, $a = i$ (l) $\dfrac{1}{z^2}$, $a = 1 + i$

6. A certain function $f(z)$ is represented by the expansion

$$\frac{1}{z^2} + \frac{1}{z^3} + \frac{1}{z^4} + \cdots$$

in $1 < |z| < \infty$. Determine the value of $f(z)$ at $z = 2i$ and at $z = i/3$.

7. A certain function $f(z)$ is represented by the expansion

$$\frac{1}{z} - \frac{1}{z^2} + \frac{1}{z^3} - \cdots$$

in $1 < |z| < \infty$. Determine the value of $f(z)$ at $z = 2$ and at $z = 1/3$.

8. Let $f(z) = \log z$ be defined by a principal value branch cut. Can f be expanded in a Laurent series about $z = 0$? Explain.

9. (*Bessel functions*) In an exercise in Section 4.6 we noted that $\exp\left[\frac{x}{2}(z - \frac{1}{z})\right]$ is the *generating function* for the Bessel functions $J_n(x)$ inasmuch as

$$e^{\frac{x}{2}(z - \frac{1}{z})} = \sum_{n=-\infty}^{\infty} J_n(x) z^n. \tag{9.1}$$

(Here, x is not the real part of z, it is an independent real variable.)

(a) Considering the analytic nature of the generating function on the left-hand side, show that (9.1) is valid in $0 < |z| < \infty$.
(b) Use (3), with C taken to be the unit circle, to derive the integral representation of $J_n(x)$,

$$J_n(x) = \frac{1}{\pi} \int_0^{\pi} \cos(n\theta - x \sin\theta)\, d\theta. \tag{9.2}$$

24.4 Classification of Singularities

Recall that if $f(z)$ is not analytic at a given point then it is singular there. Now that we have studied Laurent series we are in a position to examine the nature of singularities and to distinguish and classify them into different types.

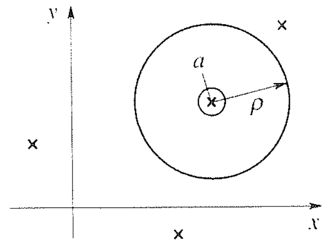


Figure 1. Isolated singular point at $z = a$.

Let $f(z)$ be singular at $z = a$. If it is analytic in an annulus $0 < |z - a| < \rho$ (for some $\rho > 0$), then $z = a$ is said to be an **isolated singular point** of f (Fig. 1); otherwise it is a **nonisolated singular point**. For example, $1/[z(z+2)]$ is singular only at $z = 0$ and $z = -2$, each of which singular point is isolated.

EXAMPLE 1. The function

$$f(z) = \frac{1}{\sin(1/z)} \tag{1}$$

is singular at $z = 1/k\pi$ $(k = \pm 1, \pm 2, \ldots)$ because $\sin(1/z) = 0$ at those points, and at $z = 0$ as well, because $\sin(1/z)$ is not even defined there (Fig. 2). Each of the former is isolated, but the singular point $z = 0$ is *not* because every annulus $0 < |z| < \rho$ inevitably contains at least one singular point (in fact, an infinite number of them) no matter how small we choose ρ. ▮

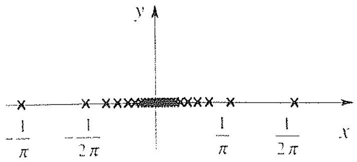


Figure 2. The singular points of $1/\sin(1/z)$.

EXAMPLE 2. The function $f(z) = \log z$, made single-valued by a branch cut, is singular at $z = 0$, its singularity being a branch point. The latter is *not* an isolated singular point because there is no annulus $0 < |z| < \rho$ throughout which f is analytic. Indeed, f is not even *defined* throughout such an annulus because of the intrusion of the branch cut. Similarly for *any* branch point singularity, such as the one at the origin for the function $z^{5/3}$, and the one at $4i$ for the function $\sqrt{z-4i}$. ▮

EXAMPLE 3. The function $f(z) = |z|^2 = x^2 + y^2$ is analytic *no*where. Thus, every point in the plane is a nonisolated singular point. ▮

In the remainder of this section we consider only isolated singular points. If $f(z)$ has an isolated singularity at $z = a$, there necessarily exists an annulus $0 < |z - a| < \rho$, for some positive number ρ, in which f admits the Laurent series representation

$$\begin{aligned} f(z) &= \sum_{n=-\infty}^{\infty} c_n (z-a)^n \\ &= \cdots + c_{-2}\frac{1}{(z-a)^2} + c_{-1}\frac{1}{z-a} + c_0 + c_1(z-a) + \cdots. \end{aligned} \tag{2}$$

If the expansion terminates on the left so that it is of the form

$$f(z) = c_{-N}\frac{1}{(z-a)^N} + c_{-N+1}\frac{1}{(z-a)^{N-1}} + \cdots, \tag{3}$$

then we categorize the singularity of f at $z = a$ as an ***N*th-order pole**. (A *first*-order pole is sometimes called a *simple* pole.) If not (i.e., if there are an infinite number of negative powers of $z - a$ present), then we categorize the singularity as an **essential singularity**.

EXAMPLE 4. The function $f(z) = 1/[z^2(1-z)]$ admits two possible expansions about the singular point $z = 0$:

$$\frac{1}{z^2(1-z)} = \frac{1}{z^2} + \frac{1}{z} + 1 + z + z^2 + \cdots \tag{4}$$

in $0 < |z| < 1$, and

$$\frac{1}{z^2(1-z)} = -\frac{1}{z^3} - \frac{1}{z^4} - \frac{1}{z^5} - \cdots \tag{5}$$

in $1 < |z| < \infty$. The former seems to indicate that f has a second-order pole at $z = 0$, and the latter seems to indicate that it has an essential singularity there. The key is to remember that the classification is based on the Laurent series (2) which is valid in an annulus in $0 < |z - a| < \rho$, that is, in an annulus with an infinitely tight inner circle! Since (5) is valid in $1 < |z| < \infty$ it is irrelevant, gives no information, insofar as the classification of the singularity at $z = 0$. Indeed, that fact is perfectly reasonable in that the region of validity is at a nonzero distance from the point in question. On the other hand, (4) is valid in $0 < |z| < 1$, hence right up to the point in question, and reveals that f has a second-order pole at $z = 0$. ▮

EXAMPLE 5. We saw, in Example 4 of Section 24.3, that

$$f(z) = e^{1/z} = 1 + \frac{1}{z} + \frac{1}{2!}\frac{1}{z^2} + \frac{1}{3!}\frac{1}{z^3} + \cdots \tag{6}$$

in $0 < |z| < \infty$. Since (6) contains an infinite number of negative powers, f has an essential singularity at $z = 0$. ▮

To proceed further it is useful to consider the vanishing of a function at a given point. We say that $f(z)$ has a **zero** at $z = a$ if $f(a) = 0$. Suppose, further, that $f(z)$ is analytic at a so that it admits a Taylor series represantation

$$f(z) = f(a) + f'(a)(z-a) + \frac{f''(a)}{2!}(z-a)^2 + \cdots \tag{7}$$

in some neighborhood of a. If $f(a) = f'(a) = \cdots = f^{(k-1)}(a) = 0$, and $f^{(k)}(a) \neq 0$, then we say that the zero is of order k. Thus, if f has a ***k*th-order zero** at a then

$$\begin{aligned} f(z) &= \frac{f^{(k)}(a)}{k!}(z-a)^k + \frac{f^{(k+1)}(a)}{(k+1)!}(z-a)^{k+1} + \cdots \\ &\sim \frac{f^{(k)}(a)}{k!}(z-a)^k \end{aligned} \tag{8}$$

as $z \to a$, where $f^{(k)}(a) \neq 0$. For example, $\sin z$ has a first-order zero at $z = 0$, $1 - \cos z$ has a second-order zero there, $\sin^3 z$ and $\sin(z^3)$ have third-order zeros there, and so on. If $f(a) \neq 0$ so that f does not have a zero at a, it will nevertheless be convenient to say that f has a **zeroth-order zero** there.

THEOREM 24.4.1 *Nth-Order Pole*
If $p(z)$ and $q(z)$ have zeros of order P and Q, respectively, at $z = a$, then $f(z) = p(z)/q(z)$ has a pole of order $N = Q - P$ there if $Q > P$, and is analytic there if $Q \leq P$.

EXAMPLE 6. Locate and classify all singularities of

$$f(z) = \frac{(\pi - z)(z^4 - 3z^2)}{\sin^2 z}. \tag{9}$$

Candidates for singular points are the zeros of the denominator, $z = n\pi$ ($n = 0, \pm 1, \pm 2, \ldots$). Of these, $z = 0$ and $z = \pi$ also happen to be zeros of the numerator so they need to be considered separately. At $z = 0$, $p(z) = (\pi - z)(z^4 - 3z^2) = -3\pi z^2 + \cdots$ has a second-order zero, and $q = \sin^2 z = z^2 + \cdots$ has a second-order zero, so $P = Q = 2$ and hence (by Theorem 24.4.1) f is analytic there.

At $z = \pi$, $p(z) = (3\pi^2 - \pi^4)(z - \pi) + \cdots$ has a first order zero. Since

$$\sin z = \sin\pi + (\cos\pi)(z-\pi) + \cdots = -(z-\pi) + \cdots, \tag{10}$$

it follows that $q(z) = \sin^2 z = (z-\pi)^2 + \cdots$ has a second-order zero. Thus, $N = Q - P = 2 - 1 = 1$ so f has a first-order pole at π.

At $z = n\pi$ for $n \neq 0$ and $n \neq 1$, $p(z)$ has a zeroth-order zero so $P = 0$. And since

$$\sin z = \sin n\pi + (\cos n\pi)(z - n\pi) + \cdots = (-1)^n(z - n\pi) + \cdots, \tag{11}$$

we see that $q(z) = \sin^2 z = (z - n\pi)^2 + \cdots$ has a second-order zero so $N = Q - P = 2 - 0 = 2$, and f has a second-order pole at each of those points. ▮

Having distinguished isolated singularities as poles or essential singularities, it is of interest to see how $f(z)$ behaves near such points.

Suppose that $f(z)$ has a pole at $z = a$, say a second-order pole for definiteness. Then

$$f(z) = \frac{c_{-2}}{(z-a)^2} + \frac{c_{-1}}{z-a} + c_0 + c_1(z-a) + \cdots \qquad (c_{-2} \neq 0) \tag{12}$$

holds in $0 < |z-a| < \rho$ for some ρ. Since $c_0 + c_1(z-a) + \cdots$ is a convergent power series in $|z-a| < \rho$, its sum, say $g(z)$, is analytic in $|z-a| < \rho$, and we can express (12) more compactly as

$$f(z) = \frac{c_{-2}}{(z-a)^2} + \frac{c_{-1}}{z-a} + g(z). \tag{13}$$

Since $g(z)$ is analytic at $z = a$ it is surely continuous there so

$$(z-a)^2 f(z) = c_{-2} + c_{-1}(z-a) + g(z)(z-a)^2 \to c_{-2}$$

as $z \to a$. Hence, $|(z-a)^2 f(z)| = |z-a|^2\,|f(z)| \to |c_{-2}| \neq 0$, and since $|z-a|^2 \to 0$ as $z \to a$, it follows that $|f(z)| \to \infty$ as $z \to a$ (Fig. 3). Similarly for a pole of *any* order so we can state the following:

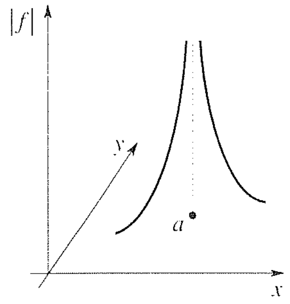


Figure 3. Growth of $|f|$ as $z \to a$.

THEOREM 24.4.2 *Behavior Near Pole*
If $f(z)$ has a pole at $z = a$, then $|f(z)| \to \infty$ as $z \to a$.

Thus, one says that $f(z)$ "blows up" as z approaches a (from any direction).

The behavior near an essential singularity is more subtle. To illustrate, consider $f(z) = \exp(-1/z^2)$, which has an essential singularity at $z = 0$, as is evident from the Laurent expansion

$$e^{-1/z^2} = 1 - \frac{1}{z^2} + \frac{1}{2!}\frac{1}{z^4} + \frac{1}{3!}\frac{1}{z^6} + \cdots, \tag{14}$$

valid in $0 < |z| < \infty$. Now, if $z \to 0$ along the real axis, then $\exp(-1/z^2) = \exp(-1/x^2) \to 0$ as $x \to 0$ from the left or right, but if $z \to 0$ along the imaginary axis, then $\exp(-1/z^2) = \exp[-1/(iy)^2] = \exp(1/y^2) \to \infty$ as $y \to 0$ from above or below.

In fact, *Emile Picard* (1856–1941) proved the following:

THEOREM 24.4.3 *Behavior Near Essential Singularity; Picard's Theorem*
Let $f(z)$ have an essential singularity at $z = a$. Then $f(z)$ takes on *every* complex number as a value, with at most one exception, and it does so an infinite number of times within any given neighborhood of a.

Let us illustrate Picard's theorem with an example.

EXAMPLE 7. Recall from Example 5 that $e^{1/z}$ has an essential singularity at $z = 0$. Now, $e^{1/z}$ does not equal 0 for any z so 0 must be the "exceptional value" referred to in the theorem, in this case. Let c be any given complex number other than this exceptional value; $c \neq 0$. Then $e^{1/z} = c$ gives $1/z = \log c = \ln|c| + i(\theta_0 + 2n\pi)$, where θ_0 is the principal argument of c. Thus, $e^{1/z}$ takes on the value c at the points

$$z_n = \frac{1}{\ln|c| + i(\theta_0 + 2n\pi)} \tag{15}$$

for $n = 0, \pm 1, \pm 2, \dots$, and (because we can choose n as large as we like) an infinite number of these points are to be found in every neighborhood $|z| < \epsilon$, no matter how small we choose ϵ. ▮

One final idea. Recall from Section 22.3 that when we wish to include the point at infinity in the z plane we speak of the *extended* z plane. Thus, when we say that $f(z) = e^z$, say, is analytic everywhere in the z plane, it is to be understood that we are not including the point at infinity. If we do wish to examine e^z in the extended z plane, we also need to examine it at infinity. To do so, we make the change of variables $z = 1/t$ so $z = \infty$ corresponds to $t = 0$, and then examine $e^z = e^{1/t}$ at $t = 0$. In this case $e^{1/t}$ has an essential singularity at $t = 0$ so we say that e^z has an essential singularity at $z = \infty$. Thus, e^z is analytic everywhere in the extended z plane except at $z = \infty$, where it has an essential singularity. Similarly,

$$f(z) = \frac{1}{z^2} + 3 + z - 4z^3 \tag{16}$$

has a second-order pole at $z = 0$ and a third-order pole at infinity because

$$f\left(\frac{1}{t}\right) = t^2 + 3 + \frac{1}{t} - \frac{4}{t^3} \tag{17}$$

has a third-order pole at $t = 0$.

Closure. Distinguishing isolated singularities from nonisolated ones, we limit our attention to the former. If $f(z)$ has an isolated singular point at $z = a$, then there is necessarily an annulus $0 < |z - a| < \rho$, for some ρ, within which f admits a Laurent series representation. If that series has an infinite number of negative powers of $z - a$, then the singularity of f at a is called an essential singularity. If instead there are only a finite number of negative powers, the degree of the most negative power being $-N$, then the singularity is called an Nth-order pole. In making that assessment, it is critical to remember that the Laurent series under consideration must hold in an infinitely tight annulus. Remember that branch point singularities are always nonisolated and therefore fall outside our discussion of poles and essential singularities.

We find that the behavior of f near a pole is quite different from its behavior near an essential singularity. Specifically, $|f(z)| \to \infty$ as z approaches a pole from any direction, whereas if f has an essential singularity at a, then the limit of f as $z \to a$ depends upon the direction of approach.

EXERCISES 24.4

1. Prove Theorem 24.4.1.

2. Determine the location of the zeros (if any) of order 1 or higher, and their order, for each given function.

(a) $z^2 - z$
(b) $e^z - 1$
(c) $z \sin z$
(d) $z \cos^2 z$
(e) $(z^2 + 1)^3$
(f) $2 + e^z$
(g) $z^2 + z + 1$
(h) $1 - z^4$

3. Determine all singular points, in the finite z plane, of the following functions. If isolated, classify them further as Nth-order poles or essential singularities.

(a) $\dfrac{e^z - 1}{z^3}$
(b) $\dfrac{z^2}{e^z - 1}$
(c) $\dfrac{1}{z^3 - 1}$
(d) $\dfrac{1}{1 + 1/(1 + z)}$
(e) $\dfrac{1}{\sinh z}$
(f) $\dfrac{1}{\cosh z}$
(g) $\dfrac{z}{\sin^3 z}$
(h) $\sin \dfrac{1}{z}$
(i) $\cos \dfrac{1}{z}$
(j) $\sinh \dfrac{1}{z}$
(k) $\cosh \dfrac{1}{z}$
(l) $\sin \dfrac{1}{z^3}$
(m) e^{1/z^3}
(n) $4 - \dfrac{z}{(z-1)^2}$
(o) $\dfrac{1}{e^z}$
(p) $\tan (z^2)$
(q) $\tan \dfrac{1}{z^2}$
(r) $\dfrac{1}{e^{z^2}}$
(s) $\dfrac{1}{\sin (z - 2)}$
(t) $\dfrac{1}{\sin \dfrac{1}{z}}$

4. (a)–(t) Is the function given in the corresponding part of Exercise 3 analytic at infinity? If not, classify its singularity there.

5. A function $f(z)$ is represented in a certain annulus by the given Laurent series. Classify the singularity (if any) of f at $z = 0$.

(a) $\dfrac{1}{z^2} + \dfrac{1}{z^3} + \dfrac{1}{z^4} + \cdots$
(b) $\dfrac{1}{2z} - \dfrac{1}{(2z)^2} + \dfrac{1}{(2z)^3} - \cdots$
(c) $\dfrac{2}{z^4} + 3z^4$
(d) $\dfrac{1}{2!\,z^2} + \dfrac{1}{3!\,z^3} + \dfrac{1}{4!\,z^4} + \cdots$
(e) $\dfrac{1}{z^5}\left[1 - \left(\dfrac{2}{z^3}\right) + \left(\dfrac{2}{z^3}\right)^2 - \left(\dfrac{2}{z^3}\right)^3 + \cdots\right]$

6. Give a critical assessment of the following claim: $f(z) = 1/e^z$ has an essential singularity at the origin because

$$\frac{1}{e^z} = \frac{1}{1 + z + \frac{z^2}{2!} + \cdots}$$

has an infinite number of negative powers of z.

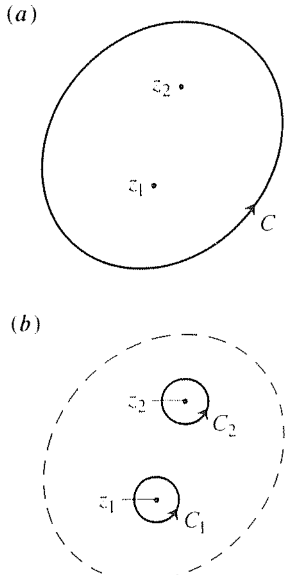


Figure 1. Deformation of contour.

24.5 Residue Theorem

In this final section we derive the powerful residue theorem for the evaluation of complex integrals (and real integrals as well) and give numerous applications.

24.5.1. Residue theorem. Consider the contour integral

$$I = \oint_C f(z)\,dz, \tag{1}$$

where f has a finite number of singular points within C, and each of them is isolated. We do not permit singularities on the contour C, and whether or not f has singular points outside C will not be relevant.

For definiteness, suppose that f has two isolated singular points within C, z_1 and z_2, as shown in Fig. 1a. Using Cauchy's theorem to deform the contour into C_1 and C_2 (Fig. 1b), we have

$$I = \oint_C f(z)\,dz = \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz. \tag{2}$$

To evaluate the latter two integrals, which we denote as I_1 and I_2, respectively, expand f in Laurent series

$$f(z) = \sum_{n=-\infty}^{\infty} c_n^{(1)}(z - z_1)^n \qquad \text{in } 0 < |z - z_1| < \rho_1 \tag{3a}$$

and

$$f(z) = \sum_{n=-\infty}^{\infty} c_n^{(2)}(z - z_2)^n \qquad \text{in } 0 < |z - z_2| < \rho_2, \tag{3b}$$

and be sure that C_1 is small enough to fit entirely within the $0 < |z - z_1| < \rho_1$ annulus, and that C_2 is small enough to fit entirely within the $0 < |z - z_2| < \rho_2$ annulus. Then (3a) holds on C_1 and (3b) on C_2, so we can re-express (2) as

$$I = \oint_{C_1} \sum_{n=-\infty}^{\infty} c_n^{(1)}(z - z_1)^n\,dz + \oint_{C_2} \sum_{n=-\infty}^{\infty} c_n^{(2)}(z - z_2)^n\,dz$$

$$
\begin{aligned}
&= \sum_{n=-\infty}^{\infty} c_n^{(1)} \oint_{C_1} (z - z_1)^n \, dz + \sum_{n=-\infty}^{\infty} c_n^{(2)} \oint_{C_2} (z - z_2)^n \, dz \\
&= 2\pi i \, c_{-1}^{(1)} + 2\pi i \, c_{-1}^{(2)}, \qquad (4)
\end{aligned}
$$

where the last step follows our "important little integral" (Example 2 in Section 24.3). Understand that (3a) and (3b) are two different Laurent series – of the same function, $f(z)$, but about z_1 and z_2, respectively. Hence, their coefficients are in general different, and we denote them as $c_n^{(1)}$ and $c_n^{(2)}$, respectively.

Observe that of the infinite number of terms in the C_1 integral, only the $n = -1$ term survives and contributes to the answer, its contribution being $2\pi i \, c_{-1}^{(1)}$; similarly for the C_2 integral. Thus, the surviving coeficients in (4), $c_{-1}^{(1)}$ and $c_{-1}^{(2)}$, are called the **residues** of $f(z)$ at z_1 and z_2, and we have found in (4) that $I = \oint_C f(z)\, dz$ is equal to $2\pi i$ times the sum of the residues.

The foregoing derivation is attractive in its directness, but it does beg justification of the termwise integration of the two Laurent series, expressed by the second equality in (4). Though that step can indeed be justified, let us sidestep the issue altogether and simply use equation (3) in Theorem 24.3.1. With $n = -1$, that equation gives, immediately,

$$
\oint_{C_1} f(z)\, dz = 2\pi i \, c_{-1}^{(1)} \qquad \text{and} \qquad \oint_{C_2} f(z)\, dz = 2\pi i \, c_{-1}^{(2)}, \qquad (5)
$$

in agreement with the result obtained above.

Surely the same method applies if C contains any finite number of isolated singular points so we can state the residue theorem:

THEOREM 24.5.1 *Residue Theorem*
Let C be a piecewise smooth simple closed curve oriented counterclockwise, and let $f(z)$ be analytic inside and on C except at finitely many isolated points $z_1, \ldots, z_k$ within C. If $c_{-1}^{(j)}$ denotes the residue of f at z_j, then

$$
\boxed{\oint_C f(z)\, dz = 2\pi i \sum_{j=1}^{k} c_{-1}^{(j)}.} \qquad (6)
$$

That is, the integral equals $2\pi i$ times the sum of the residues of f within C.

Again, by the residue of f at z_j we mean the $c_{-1}^{(j)}$ coefficient in the Laurent expansion

$$
f(z) = \sum_{n=-\infty}^{\infty} c_n^{(j)} (z - z_j)^n \qquad (7)
$$

of f about z_j, in some annulus $0 < |z - z_j| < \rho_j$.

EXAMPLE 1. Evaluate

$$I = \oint_C z^4 \sin\frac{1}{z}\,dz, \tag{8}$$

where C is the circle $|z| = 1$, counterclockwise. There is only one singular point, the essential singular point at $z = 0$. The relevant Laurent series is

$$\begin{aligned} z^4 \sin\frac{1}{z} &= z^4\left(\frac{1}{z} - \frac{1}{3!\,z^3} + \frac{1}{5!\,z^5} - \frac{1}{7!\,z^7} + \cdots\right) \\ &= z^3 - \frac{z}{3!} + \frac{1}{5!\,z} - \frac{1}{7!\,z^3} + \cdots, \qquad (0 < |z| < \infty) \end{aligned} \tag{9}$$

so the residue is $1/5! = 1/120$, and $I = 2\pi i(1/120) = \pi i/60$.

COMMENT 1. If C were clockwise instead, then we would have $I = -\pi i/60$.

COMMENT 2. If the integrand were $z^3 \sin\frac{1}{z}$ instead, then in place of (9) we would have

$$z^3 \sin\frac{1}{z} = z^2 - \frac{1}{3!} + \frac{1}{5!\,z^2} - \frac{1}{7!\,z^4} + \cdots. \tag{10}$$

It would be incorrect to say "there is no residue" because there is no $1/z$ term. Rather, the coefficient of the $1/z$ term happens to be 0 so the residue is 0 and $I = (2\pi i)(0) = 0$. ▮

In Example 1 it was easy to write out the desired Laurent series, and hence to pick out the residue (as the coefficient of $1/z$). However, the beauty of the residue theorem is that we don't need the entire Laurent series, all we need is one coefficient in it, the c_{-1} residue. Thus, let us develop a method for evaluating the residue without having to generate the entire Laurent series.

24.5.2. Calculating residues. To begin, suppose that $f(z)$ has a first-order pole at a so that

$$f(z) = c_{-1}\frac{1}{z-a} + c_0 + c_1(z-a) + \cdots \tag{11}$$

in $0 < |z - a| < \rho$ for some ρ. Then

$$(z-a)f(z) = c_{-1} + c_0(z-a) + c_1(z-a)^2 + \cdots, \tag{12}$$

and letting $z \to a$ in both sides gives

$$c_{-1} = \lim_{z\to a}[(z-a)f(z)]. \tag{13}$$

Next, suppose that $f(z)$ has an Nth-order pole at $z = a$ so

$$f(z) = c_{-N}\frac{1}{(z-a)^N} + c_{-N+1}\frac{1}{(z-a)^{N-1}} + \cdots. \tag{14}$$

Then

$$(z-a)^N f(z) = c_{-N} + c_{-N+1}(z-a) + c_{-N+2}(z-a)^2 + \cdots. \tag{15}$$

Unfortunately, letting $z \to a$ gives c_{-N} rather than the desired coefficient c_{-1}. However, the right side of (15) is the Taylor series of $(z-a)^N f(z)$ so it follows that the coefficient c_{-N+j} of $(z-a)^j$ is the jth derivative of $(z-a)^N f(z)$ evaluated at a and divided by $j!$. Since c_{-N+j} becomes c_{-1} when $j = N-1$, it follows that

$$\boxed{c_{-1} = \frac{1}{(N-1)!}\lim_{z\to a}\left\{\frac{d^{N-1}}{dz^{N-1}}\left[(z-a)^N f(z)\right]\right\}.} \tag{16}$$

To use (16) we need to know the order N of the pole; *if the singularity is an essential singularity, then (16) does not apply.*

EXAMPLE 2. Evaluate all residues of

$$f(z) = \frac{1}{(z+4)(z-1)^3}. \tag{17}$$

The denominator has first- and third-order zeros at -4 and 1, respectively, and the numerator has zeroth-order zeros at those points so f has a first-order pole ($N = 1$) at $z = -4$ and a third-order pole ($N = 3$) at $z = 1$. Thus, (16) gives

$$\begin{aligned}\operatorname*{Res}_{z=-4} f &= \frac{1}{0!}\lim_{z\to-4}\left[(z+4)\frac{1}{(z+4)(z-1)^3}\right]\\ &= \lim_{z\to-4}\frac{1}{(z-1)^3} = -\frac{1}{125},\end{aligned} \tag{18}$$

and

$$\begin{aligned}\operatorname*{Res}_{z=1} f &= \frac{1}{2!}\lim_{z\to1}\left\{\frac{d^2}{dz^2}\left[(z-1)^3\frac{1}{(z+4)(z-1)^3}\right]\right\}\\ &= \frac{1}{125}\end{aligned} \tag{19}$$

as the desired residues. ▮

24.5.3. Applications of the residue theorem. Let us consider several applications.

EXAMPLE 3. Evaluate $I = \oint_C f(z)\,dz$, where $f(z)$ is given by (17) and C is counterclockwise. If C encloses both singular points, then the residue theorem gives

$I = 2\pi i(-\frac{1}{125} + \frac{1}{125}) = 0$; if C encloses only one of them, say the one at $z = 1$, then $I = 2\pi i(\frac{1}{125}) = 2\pi i/125$; and if C encloses neither of them, then $I = 0$. ∎

Actually, most practical applications of the residue theorem are to *real* integrals. How can that be? We shall see, in the following examples.

EXAMPLE 4. Evaluate the real integral

$$I = \int_{-\infty}^{\infty} \frac{dx}{x^2+1}. \tag{20}$$

That (20) is a "real integral" is no problem because we can just as well write

$$I = \int_C \frac{dz}{z^2+1},$$

where C is the path from $-\infty$ to $+\infty$ along the x axis. What *is* a problem, however, is that to apply the residue theorem we need C to be a *closed* path. Thus, we will consider, in place of I, the contour integral

$$J = \int_C \frac{dz}{z^2+1}, \tag{21}$$

where C is the closed path shown in Fig. 2, not from $-\infty$ to ∞ but from $-R$ to $+R$ and then closed with a semicircle.

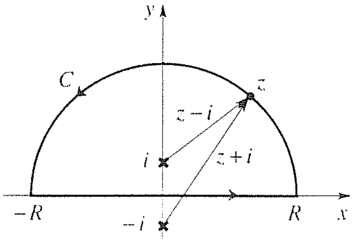


Figure 2. The closed contour C in (21).

The integrand has first-order poles at $z = +i$ and $-i$, and is analytic elsewhere. Of these, only the pole at $+i$ is within C. Thus, on the one hand, the residue theorem gives

$$J = 2\pi i \operatorname*{Res}_{z=i} f = 2\pi i \lim_{z\to i} \left[(z-i)\frac{1}{(z+i)(z-i)}\right] = \pi \tag{22}$$

(provided that $R > 1$, so that $z = i$ is within C). On the other hand,

$$J = \int_{-R}^{R} \frac{dx}{x^2+1} + \int_{C_R} \frac{dz}{z^2+1}, \tag{23}$$

where C_R denotes the semicircular part of C. Equating these two results gives

$$\pi = \int_{-R}^{R} \frac{dx}{x^2+1} + \int_{C_R} \frac{dz}{z^2+1}. \tag{24}$$

Now, (24) holds for all R's (greater than 1) so it must hold as $R \to \infty$. Taking that limit, (24) becomes*

$$\pi = \int_{-\infty}^{\infty} \frac{dx}{x^2+1} + \lim_{R\to\infty} \int_{C_R} \frac{dz}{z^2+1}. \tag{25}$$

*Recall from Section 4.5 that the singular integral $\int_a^\infty f(x)\,dx$ is defined as the limit of $\int_a^B f(x)\,dx$ as $B \to \infty$. The case $\int_{-\infty}^{\infty} f(x)\,dx$ was not covered there. The latter integral is defined as the limit of $\int_{-A}^{B} f(x)\,dx$ as A and B tend to infinity *independently*. *If* that limit does indeed exist, then there is no harm in letting $A = B$ and writing $\int_{-\infty}^{\infty} f(x)\,dx$ as the limit of $\int_{-A}^{A} f(x)\,dx$ as $A \to \infty$, as occurs in our passage from (24) to (25).

To evaluate the last term in (25) we use the ML bound. From Fig. 2 we see that as z traverses C_R the $z - i$ "vector" is smallest when $z = Ri$, so its smallest magnitude is $R - 1$. Further, the $z + i$ "vector" is smallest when $z = \pm R$ so its smallest magnitude is $\sqrt{R^2 + 1}$. Thus,

$$\left| \frac{1}{z^2 + 1} \right| = \frac{1}{|z - i|\,|z + i|} \leq \frac{1}{(R - 1)\sqrt{R^2 + 1}} \tag{26}$$

for all z on C_R, so the latter can be used as M. And since the length of C_R is πR, we have from the ML bound,

$$\left| \int_{C_R} \frac{dz}{z^2 + 1} \right| \leq \frac{1}{(R - 1)\sqrt{R^2 + 1}} \pi R \sim \frac{\pi}{R}, \tag{27}$$

which tends to zero as $R \to \infty$. Thus, (25) becomes $\pi = I + 0$, so $I = \pi$.

COMMENT 1. The method employed in this example is fairly general so let us review it. Given an integral I from $-\infty$ to ∞ on a real axis, we considered instead a contour integral J on a closed contour in a z plane. On the one hand, we could evaluate J by the residue theorem and, on the other hand, J could be expressed as the desired integral I plus a computable integral: specifically, the upper semicircle contribution was shown to tend to zero as $R \to \infty$.

If we were given the semi-infinite integral

$$I = \int_0^\infty \frac{dx}{x^2 + 1}$$

instead, we could use the fact that the integrand $1/(x^2 + 1)$ is an even function to re-express I as

$$I = \frac{1}{2} \int_{-\infty}^\infty \frac{dx}{x^2 + 1},$$

and then proceed as before.

COMMENT 2. Observe that, alternatively, we could have closed the contour below, as in Fig. 3. In that case

$$\begin{aligned} J &= -2\pi i \operatorname*{Res}_{z=-i} f \qquad \text{(the minus sign because } C \text{ is clockwise)} \\ &= -2\pi i \lim_{z \to -i} \left[(z + i) \frac{1}{(z + i)(z - i)} \right] = \frac{-2\pi i}{-2i} = \pi, \end{aligned}$$

so we obtain (25) once again, where this time C_R is the lower semicircle. In the same manner as before, we can show that the C_R integral tends to zero as $R \to \infty$ so we obtain the same final result, $I = \pi$. ▮

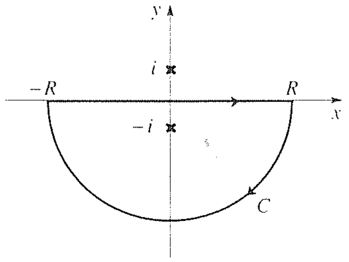


Figure 3. Closing C below.

It is true that the integral (20) was simple, and could have been evaluated by a trigonometric change of variables, giving $I = \tan^{-1} x|_{-\infty}^{\infty} = \pi$, but our chief purpose was to develop the contour integration solution method. That method applies equally well to any convergent integral of the form

$$I = \int_{-\infty}^\infty \frac{p(x)}{q(x)}\, dx, \tag{28}$$

where $p(x)$ and $q(x)$ are finite-degree polynomials.

EXAMPLE 5. Evaluate

$$I = \int_0^\infty \frac{\cos ax}{x^2+1}\,dx. \qquad (a > 0) \tag{29}$$

First, extend the lower integration limit back to $-\infty$ by noting that the integrand is an even function of x so that

$$I = \frac{1}{2}\int_{-\infty}^\infty \frac{\cos ax}{x^2+1}\,dx. \tag{30}$$

Next, consider the contour integral

$$J = \oint_C \frac{\cos az}{z^2+1}\,dz, \tag{31}$$

where C is shown in Fig. 2. Looking ahead, we will wish to show that the C_R integral tends to zero as $R \to \infty$. Does that step appear to bo feasible? Recall that

$$\cos az = \frac{e^{iaz}+e^{-iaz}}{2}.$$

If we consider the single point $z = Ri$ on C_R, we see that the e^{-iaz} term in $\cos az$ is $e^{-ia(Ri)} = e^{aR}$ there. This exponential growth (as $R \to \infty$) makes it appear highly unlikely that the C_R integral tends to zero as $R \to \infty$. Furthermore, if we try closing the contour below instead, as in Fig. 3, then at the point $z = -Ri$ on C_R the e^{+iaz} term in $\cos az$ is $e^{ia(-Ri)} = e^{aR}$. Thus, if the e^{-iaz} term does not cause trouble, the e^{+iaz} term does!

To overcome this difficulty, let us consider a *single* exponential in place of $\cos az$. Specifically, consider

$$J = \oint_C \frac{e^{iaz}}{z^2+1}\,dz \tag{32}$$

in place of (31), with C as shown in Fig. 2. The integrand $f(z)$ is analytic everywhere except at $z = \pm i$, where it has first-order poles. Thus, on the one hand,

$$J = 2\pi i \operatorname*{Res}_{z=i} f = 2\pi i\,\frac{e^{ia(i)}}{2i} = \pi e^{-a}, \tag{33}$$

and on the other hand,

$$J = \int_{-R}^R \frac{e^{iax}}{x^2+1}\,dx + \int_{C_R} \frac{e^{iaz}}{z^2+1}\,dz \tag{34}$$

or, with $R \to \infty$,

$$J = \int_{-\infty}^\infty \frac{\cos ax + i\sin ax}{x^2+1}\,dx + \lim_{R\to\infty}\int_{C_R} \frac{e^{iaz}}{z^2+1}\,dz. \tag{35}$$

Since $\left|e^{iaz}\right| = \left|e^{ia(x+iy)}\right| = \left|e^{iax}\right|\left|e^{-ay}\right| = e^{-ay} \le 1$ on C_R, it follows that

$$\left|\frac{e^{iaz}}{z^2+1}\right| \le \left|\frac{1}{z^2+1}\right| \le \frac{1}{(R-1)\sqrt{R^2+1}} \tag{36}$$

everywhere on C_R, where the last inequality follows from (26). Furthermore, the length of C_R is πR, so the ML bound gives

$$\left|\int_{C_R} \frac{e^{iaz}}{z^2+1}\,dz\right| \le \frac{\pi R}{(R-1)\sqrt{R^2+1}} \sim \frac{\pi}{R}, \tag{37}$$

which does tend to zero as $R \to \infty$. With this result, comparison of (33) and (35) gives

$$\pi e^{-a} = \int_{-\infty}^{\infty} \frac{\cos ax}{x^2+1}\,dx + i\int_{-\infty}^{\infty} \frac{\sin ax}{x^2+1}\,dx. \tag{38}$$

Finally, equating real parts in (38) gives

$$\int_{-\infty}^{\infty} \frac{\cos ax}{x^2+1}\,dx = \pi e^{-a}, \tag{39}$$

and hence the answer

$$I = \frac{1}{2}\int_{-\infty}^{\infty} \frac{\cos ax}{x^2+1}\,dx = \frac{\pi}{2}e^{-a}. \tag{40}$$

Of course, we can also equate imaginary parts in (38), and that step gives the "bonus" result

$$\int_{-\infty}^{\infty} \frac{\sin ax}{x^2+1}\,dx = 0, \tag{41}$$

but (41) is not very interesting since it follows from the fact that the integrand is an odd function. ▮

EXAMPLE 6. Evaluate

$$I = \int_0^{\infty} \frac{x^{1/3}}{(x+1)^2}\,dx. \tag{42}$$

To do so, consider the contour integral

$$J = \oint_C \frac{z^{1/3}}{(z+1)^2}\,dz, \tag{43}$$

where C is to be a suitably chosen contour. Before selecting C it is important to notice that although the $x^{1/3}$ appearing in (42) is single-valued (e.g., $8^{1/3} = 2$), the $z^{1/3}$ in (43) is *multi*-valued (e.g., $8^{1/3} = 2, 2e^{i2\pi/3}, 2e^{i4\pi/3}$). Having written down the $z^{1/3}$, we need to assume responsibility for defining it. As we will see, a branch cut to the right, as depicted in Fig. 4a, will be convenient.

Next, let us select the contour C shown in Fig. 4b, where the outer circle is of radius R and the inner circle is of radius ϵ. (We plan to let $R \to \infty$ and $\epsilon \to 0$.) It consists of four parts: PQ, QS, ST, and TP. Denote the counterclockwise circle QS as C_R, and the clockwise circle TP as C_ϵ.

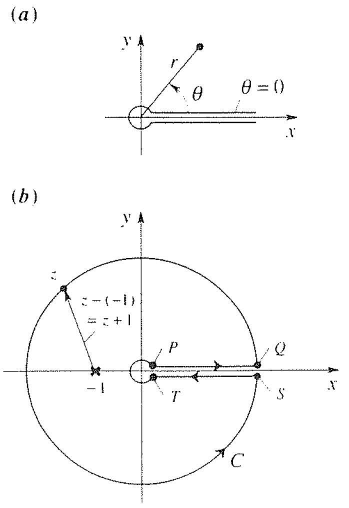


Figure 4. Branch cut for $z^{1/3}$, and contour C.

Observe carefully that on PQ we have $z = xe^{i0}$ so

$$z^{1/3} = x^{1/3}e^{i0/3} = x^{1/3}, \tag{44a}$$

whereas on ST we have $z = xe^{2\pi i}$ so

$$z^{1/3} = x^{1/3}e^{2\pi i/3}. \tag{44b}$$

And since the integrand $f(z)$ is analytic inside and on C, except for a second-order pole at $z = -1$, we have

$$\begin{aligned} 2\pi i \operatorname*{Res}_{z=-1} f(z) &= \int_\epsilon^R \frac{x^{1/3}}{(x+1)^2}\,dx + \int_{C_R} \frac{z^{1/3}}{(z+1)^2}\,dz \\ &\quad + \int_R^\epsilon \frac{x^{1/3}e^{2\pi i/3}}{(x+1)^2}\,dx + \int_{C_\epsilon} \frac{z^{1/3}}{(z+1)^2}\,dz. \end{aligned} \tag{45}$$

Applying the ML bound to the C_R integral, observe that

$$\left|\frac{z^{1/3}}{(z+1)^2}\right| = \frac{|R^{1/3}e^{i\theta/3}|}{|z+1|^2} = \frac{R^{1/3}}{|z-(-1)|^2} \le \frac{R^{1/3}}{(R-1)^2} \tag{46}$$

on C_R and that the length of C_R is $2\pi R$ so that

$$\left|\int_{C_R} \frac{z^{1/3}}{(z+1)^2}\,dz\right| \le \frac{R^{1/3}}{(R-1)^2}\,2\pi R \sim \frac{2\pi}{R^{2/3}}, \tag{47}$$

which tends to zero as $R \to \infty$. Similarly,

$$\left|\int_{C_\epsilon} \frac{z^{1/3}}{(z+1)^2}\,dz\right| \le \frac{\epsilon^{1/3}}{(1-\epsilon)^2}\,2\pi\epsilon \sim 2\pi\epsilon^{4/3}, \tag{48}$$

which tends to zero as $\epsilon \to 0$.

Furthermore,

$$\begin{aligned} \operatorname*{Res}_{z=-1} f(z) &= \frac{1}{1!}\lim_{z\to-1}\frac{d}{dz}\left[(z+1)^2\frac{z^{1/3}}{(z+1)^2}\right] = \frac{1}{3}(-1)^{-2/3} \\ &= \frac{1}{3}(1e^{\pi i})^{-2/3} = \frac{e^{-2\pi i/3}}{3}, \end{aligned} \tag{49}$$

where, in evaluating $(-1)^{-2/3}$, we have expressed $-1 = 1e^{\pi i}$, in accordance with the branch cut shown in Fig. 4a.

Thus, letting $R \to \infty$ and $\epsilon \to 0$ in (45) gives

$$\begin{aligned} 2\pi i\,\frac{e^{-2\pi i/3}}{3} &= \int_0^\infty \frac{x^{1/3}}{(x+1)^2}\,dx + 0 + \int_\infty^0 \frac{x^{1/3}e^{2\pi i/3}}{(x+1)^2}\,dx + 0 \\ &= (1 - e^{2\pi i/3})\int_0^\infty \frac{x^{1/3}}{(x+1)^2}\,dx, \end{aligned} \tag{50}$$

or,

$$I = \int_0^\infty \frac{x^{1/3}}{(x+1)^2}\,dx = \frac{2\pi i}{3}\,\frac{e^{-2\pi i/3}}{1-e^{2\pi i/3}} = \frac{2\pi}{3\sqrt{3}}. \tag{51}$$

COMMENT 1. Observe that as R tended to infinity and ϵ tended to zero the contributions from PQ, QS, ST, and TP tended either to zero or to a scalar multiple of I, so the final equation (50) was one equation in the one unknown I.

COMMENT 2. Although it was important to choose the branch cut to the right, the choice $\theta = 0$ on the top of the cut was not critical; any integer multiple of 2π would have worked just as well. ▮

From Examples 4, 5, and 6 we can see the general pattern involved in using the residue theorem to evaluate real integrals. The key is in suitably choosing the "J integral," namely, its integrand and its closed contour. We suggest choosing the same integrand as in the I integrand (with the x's changed to z's), and then modifying it only if we find that it doesn't work. For instance, in Example 5 we found that $(\cos az)/(z^2+1)$ did not work since the contribution from the semicircular part of the contour C_R did not tend to zero as $R \to \infty$, so we modified the integrand, slightly, to $e^{iaz}/(z^2+1)$. As for the contour C, the idea is to choose C so that one segment of it gives the desired I integral, and each other segment is either known or is some multiple of I. In Example 6, for instance, PQ gave I, QS went to zero, ST gave $-e^{2\pi i/3}$ times I, and TP went to zero as ϵ tended to zero and R tended to infinity.

It should be appreciated that each example given in this section is representative of an entire class of applications. Example 4, for instance, is representative of the class of integrals of the form $I = \int_{-\infty}^{\infty} p(x)\,dx/q(x)$, where p and q are polynomials.

Next, consider the class of real integrals of the form

$$I = \int_0^{2\pi} F(\cos\theta, \sin\theta)\,d\theta, \tag{52}$$

where F is a rational function of each of its two arguments. A **rational function** is the ratio of two finite-degree polynomials so $F(x,y)$ is a rational function of each of its two arguments if it is a finite linear combination of terms of the form $x^m y^n$ divided by another such finite linear combination. For example,

$$F_1(x,y) = \frac{5 - x^4 + xy^3}{x + 4y^2}$$

is a rational function of both x and y, and

$$F_2(\cos\theta, \sin\theta) = \frac{2 - \cos^2\theta\sin\theta}{1+\cos\theta}$$

and

$$F_3(\cos\theta, \sin\theta) = \frac{\sin 4\theta}{(1+\cos\theta)^2}$$

are rational functions of $\cos\theta$ and $\sin\theta$. [In the case of F_3, we need to recall the trigonometric identity

$$\sin 4\theta = \sin\theta(4\sin\theta - 8\sin^3\theta).$$

Similar identities exist for $\cos k\theta$ and $\sin k\theta$ for $k = 2, 3, 4, \dots$.]

Integrals of the form shown in (52) arise, for example, in evaluating the coefficients in Fourier series. To evaluate I, enter the change of variables

$$z = e^{i\theta}. \tag{53}$$

This change of variables converts I from an integral on a real θ axis to a contour integral in a complex z plane for, as θ varies from 0 to 2π, z undergoes one complete counterclockwise trip around the unit circle (the unit circle because $z = e^{i\theta}$ is of the form $z = re^{i\theta}$ with $r = 1$).

From (53), $dz = ie^{i\theta}\,d\theta = iz\,d\theta$ so $d\theta = dz/iz$, and

$$\begin{aligned}\cos\theta &= \frac{e^{i\theta}+e^{-i\theta}}{2} = \frac{z+z^{-1}}{2} = \frac{z^2+1}{2z},\\ \sin\theta &= \frac{e^{i\theta}-e^{-i\theta}}{2i} = \frac{z-z^{-1}}{2i} = \frac{z^2-1}{2iz},\end{aligned} \tag{54}$$

so

$$I = \int_C F\left(\frac{z+z^{-1}}{2}, \frac{z-z^{-1}}{2i}\right)\frac{dz}{iz}. \tag{55}$$

Now, we see from (54) that $\cos\theta$ and $\sin\theta$ are rational functions of z. In turn, F is a rational function of $\cos\theta$ and $\sin\theta$ so F (and hence the new integrand F/iz) is a rational function of z. Thus, the integrand in (55) is analytic everywhere except at those zeros of the denominator that lie within the contour C, that is, within the unit circle. Let us illustrate.

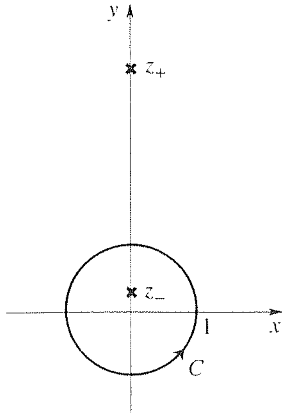


Figure 5. Contour and singular points.

EXAMPLE 7. Evaluate

$$I = \int_0^{2\pi}\frac{d\theta}{2-\sin\theta}. \tag{56}$$

With the change of variables $z = e^{i\theta}$, we obtain

$$\begin{aligned}I &= \oint_C \frac{1}{2-\dfrac{z^2-1}{2iz}}\frac{dz}{iz} = -2\oint_C \frac{dz}{z^2-4iz-1}\\ &= -2\oint_C \frac{dz}{(z-z_+)(z-z_-)} \equiv \oint_C f(z)\,dz,\end{aligned} \tag{57}$$

where $z_+ = (2+\sqrt{3})i$ and $z_- = (2-\sqrt{3})i$. Since z_+ lies outside C and z_- lies inside (Fig. 5), the residue theorem gives

$$\begin{aligned} I &= 2\pi i \operatorname*{Res}_{z=z_-} f = 2\pi i \lim_{z\to z_-}\left[(z-z_-)\frac{-2}{(z-z_+)(z-z_-)}\right] \\ &= 2\pi i\left(\frac{-2}{z_- - z_+}\right) = \frac{2\pi}{\sqrt{3}}. \end{aligned} \tag{58}$$

COMMENT 1. Remember that *the residue theorem does not apply if there are singular points on* C for the theorem requires $f(z)$ to be analytic inside and on C. For example, if we generalize (15) to the form

$$I = \int_0^{2\pi} \frac{d\theta}{a - \sin\theta}, \tag{59}$$

where a is real, we find that the residue theorem does not apply if $-1 \le a \le 1$ because in that case the singular points $\pm\sqrt{1-a^2} + ai$ fall on the contour C.

COMMENT 2. It is also to be noted that if the integral limits in (59) were 0 and π, say, the residue theorem could not have been applied because the corresponding contour C in the z plane would not have been closed; it would have been only a semicircle. ▮

As our final applications, we indicate how the residue theorem can be used to evaluate inverse Laplace and Fourier transforms. Recall that we introduced the Laplace transform of a function $f(t)$ as

$$L\{f(t)\} = F(s) = \int_0^\infty f(t)e^{-st}\,dt, \tag{60}$$

and regarded the transform variable s as a real number. For the transform $F(s)$ to *exist*–that is, for the integral in (60) to converge, we asked $f(t)$ to be of exponential order, whereby there exist real constants K, c, and T such that $|f(t)| \le Ke^{ct}$ for all $t > T$, for then we can ensure the existence of the transform $F(s)$ by asking s to exceed c.

Later, in Section 17.11, we show how to derive both the transform formula (60) and also the *Laplace inversion formula*

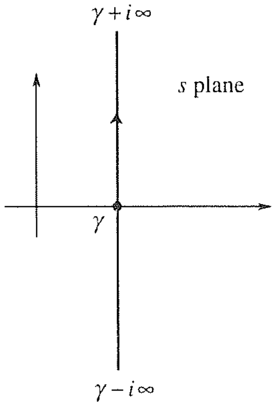


Figure 6. The contour in (61).

$$L^{-1}\{F(s)\} = f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)e^{st}\,ds. \tag{61}$$

Thus, even though it is permissible, in working with the Laplace transform in Chapter 5, to consider s as real, the inversion integral is actually along the path shown in Fig. 6 in a complex s plane. Just as we needed $s > c$ in (60) when we regarded s as real, we need $\operatorname{Re} s = \gamma > c$ now that we are considering s to be complex.

EXAMPLE 8. *Inverse Laplace Transform.* To illustrate the use of (61), let us determine the inverse of the Laplace transform

$$F(s) = \frac{1}{(s-2)^3}, \tag{62}$$

namely,

$$f(t)=\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty}\frac{e^{st}}{(s-2)^3}\,ds. \tag{63}$$

As noted above, we need γ to be sufficiently positive, that is, we need $\gamma > c$, but we don't know c yet because we don't know $f(t)$ yet! We assert, without proof, that γ *will be sufficiently positive if the line of integration (from $\gamma - i\infty$ to $\gamma + i\infty$) is placed to the right of all singularities of* $F(s)$. In the present example $F(s)$ is singular only at $s = 2$, where it has a third-order pole, so we choose $\gamma > 2$. To obtain a closed contour, shall we close on the left or on the right? Let us close on the left as shown in Fig. 7, with a semicircle of radius R centered at $s = \gamma$. Thus, consider the contour integral

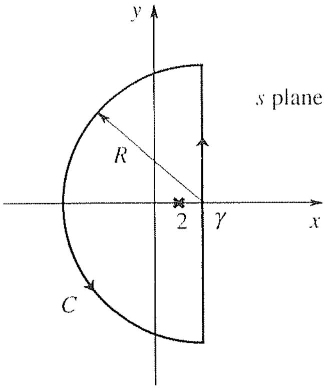


Figure 7. Laplace inversion contour.

$$I=\frac{1}{2\pi i}\oint_C\frac{e^{st}}{(s-2)^3}\,ds, \tag{64}$$

where C is shown in the figure.

On the one hand,

$$J=2\pi i\operatorname*{Res}_{s=2}\frac{e^{st}}{2\pi i(s-2)^3}=\frac{1}{2}t^2e^{2t}, \tag{65}$$

and on the other hand,

$$J=\frac{1}{2\pi i}\int_{\gamma-iR}^{\gamma+iR}\frac{e^{st}}{(s-2)^3}\,ds+\frac{1}{2\pi i}\int_{C_R}\frac{e^{st}}{(s-2)^3}\,ds, \tag{66}$$

where C_R is the semicircular part of C. Equating these results, letting $R \to \infty$ and recalling (63) gives

$$\frac{1}{2}t^2e^{2t}=f(t)+\frac{1}{2\pi i}\lim_{R\to\infty}\int_{C_R}\frac{e^{st}}{(s-2)^3}\,ds. \tag{67}$$

Now,

$$\left|e^{st}\right|=\left|e^{(x+iy)t}\right|=\left|e^{xt}\right|\left|e^{iyt}\right|=e^{xt}\le e^{\gamma t}, \tag{68}$$

and

$$|s-2|\ge R-2 \tag{69}$$

for all points s on C_R so

$$\left|\frac{e^{st}}{(s-2)^3}\right|\le\frac{e^{\gamma t}}{(R-2)^3} \tag{70}$$

on C_R. Hence the ML bound gives

$$\left|\int_{C_R}\frac{e^{st}}{(s-2)^3}\,ds\right|\le\frac{e^{\gamma t}}{(R-2)^3}\pi R\sim\frac{\pi e^{\gamma t}}{R^2}, \tag{71}$$

which tends to zero as $R \to \infty$ (with t held constant). Thus, (67) gives

$$f(t)=\frac{1}{2}t^2e^{2t}, \tag{72}$$

which result agrees with the corresponding entry in Appendix C.

COMMENT 1. With hindsight, we can now see from (72) that $|f(t)| \leq Ke^{ct}$ for some constant K and t sufficiently large, if c is any number greater than 2. Thus, we need $\gamma > 2$, and that is precisely in accord with the italicized assertion below equation (63).

COMMENT 2. You may recall, from Section 17.11, that the inversion integral in (61) gives $f(t)$ for $t > 0$; for $t < 0$ it gives zero:

$$L^{-1}\{F(s)\} = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} F(s)e^{st}\,ds = \begin{cases} f(t), & t > 0 \\ 0, & t < 0. \end{cases} \tag{73}$$

Thus far, in this example, we have understood t to be positive, which is of course the interval of interest in applications. However, it will be instructive to show that we do indeed obtain zero, in this example, for $t < 0$ in accordance with (73). Specifically, observe that the $e^{xt} \leq e^{\gamma t}$ inequality in (68) holds on C_R only if $t > 0$. Thus, the foregoing analysis breaks down for $t < 0$, and we need to close the contour on the right instead as shown in Fig. 8. We leave it for the exercises to show that with this contour used in (64) we do obtain $L^{-1}\{F(s)\} = 0$.

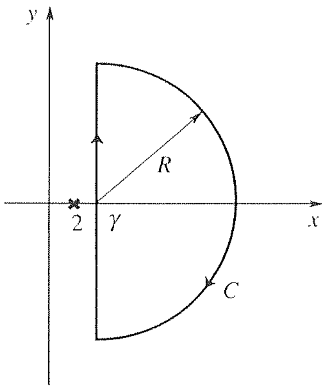


Figure 8. Inversion contour for $t < 0$.

COMMENT 3. In summary, the idea is to keep the line from $\gamma - i\infty$ to $\gamma + i\infty$ to the right of all singularities of $F(s)$ and to close the contour on the left for $t > 0$ and on the right for $t < 0$. Of course, the case $t < 0$ need not be carried out because the result will always be zero. ▮

Similarly, to find the inverse of a Fourier transform $\hat{f}(\omega)$ we use the *Fourier inversion formula*

$$F^{-1}\{\hat{f}(\omega)\} = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}\,d\omega. \tag{74}$$

EXAMPLE 9. *Inverse Fourier Transform.* Determine the inverse of the Fourier transform

$$\hat{f}(\omega) = \frac{1}{\omega^2+1}, \tag{75}$$

namely,

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{\omega^2+1}\,e^{i\omega x}\,d\omega. \tag{76}$$

To do so, consider the contour integral

$$J = \frac{1}{2\pi}\oint_C \frac{e^{ix\omega}}{\omega^2+1}\,d\omega \tag{77}$$

in the complex ω plane, where C is the closed contour shown in Fig. 9a for $x > 0$ and the closed contour shown in Fig. 9b for $x < 0$. (We close the contour above for $x > 0$ and below for $x < 0$ so that in each case the integral on the semicircular part of C will tend to zero as $R \to \infty$.)

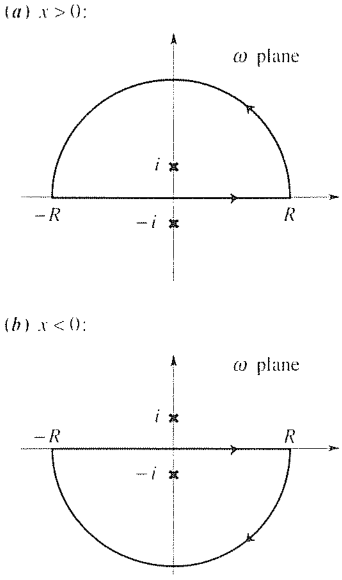


Figure 9. Contour C for (77).

First, consider the case $x > 0$. On the one hand,

$$J = 2\pi i \operatorname*{Res}_{\omega = i} \left(\frac{1}{2\pi} \frac{e^{ix\omega}}{\omega^2 + 1} \right) = 2\pi i \frac{1}{2\pi} \frac{e^{ix(i)}}{2i} = \frac{e^{-x}}{2}, \tag{78}$$

and on the other hand,

$$J = \frac{1}{2\pi} \int_{-R}^{R} \frac{e^{ix\omega}}{\omega^2 + 1} \, d\omega + \frac{1}{2\pi} \int_{C_R} \frac{e^{ix\omega}}{\omega^2 + 1} \, d\omega, \tag{79}$$

where C_R denotes the semicircular part of C in Fig. 9a. As $R \to \infty$ the first integral in (79) tends to $f(x)$ [recall (76)] and the second integral tends to zero. [The bounding process is the same as in (36) and (37), with z changed to ω and a changed to x, so we will not repeat it here.] Thus, letting $R \to \infty$ in (79) gives

$$J = f(x) + 0, \tag{80}$$

and comparison of (78) and (80) shows that

$$f(x) = \frac{e^{-x}}{2}. \qquad (x > 0) \tag{81}$$

Next, consider $x < 0$, remembering that this time we close the contour below (Fig. 9b). On the one hand,

$$J = -2\pi i \operatorname*{Res}_{\omega = -i} \left(\frac{1}{2\pi} \frac{e^{ix\omega}}{\omega^2 + 1} \right) = -2\pi i \frac{1}{2\pi} \frac{e^{ix(-i)}}{-2i} = \frac{e^{x}}{2}, \tag{82}$$

the first minus sign being necessary because this time C is clockwise, and on the other hand

$$J = \frac{1}{2\pi} \int_{-R}^{R} \frac{e^{ix\omega}}{\omega^2 + 1} \, d\omega + \frac{1}{2\pi} \int_{C_R} \frac{e^{ix\omega}}{\omega^2 + 1} \, d\omega, \tag{83}$$

where C_R is the semicircular part of C in Fig. 9b. Letting $R \to \infty$ once again, (83) gives

$$J = f(x) + 0, \tag{84}$$

and comparison of (82) and (84) shows that

$$f(x) = \frac{e^{x}}{2}. \qquad (x < 0) \tag{85}$$

Finally, (81) and (85) give

$$f(x) = \frac{e^{-|x|}}{2} \tag{86}$$

for all x, which result agrees with entry 4 of Appendix D for the choice $a = 1$. ▮

Closure. This section underscores the power of analytic function theory by revealing that the integral around a closed path is simply $2\pi i$ times the sum of the residues contained within the path. (For a precise statement, see Theorem 24.5.1.) And it is not difficult to calculate residues, even if the integrand is quite unwieldy.

Besides closed loop integrals in the complex plane, we find that the residue theorem can be used to evaluate a wide variety of integrals on the real axis, and inverse Laplace transforms as well. A key step is in setting up the "J integral" – that is, in choosing its integrand and the closed contour C. The integrand is often, though not always, the same as the original integrand but with the original real integration variable replaced by the complex variable z. As an example of this rule, compare (21) with (20). As an example of an exception to the rule, compare (32) with (30). If the contour is not closed we need to close it, so we can apply the residue theorem.

A basic question arises, which we have not yet addressed: Is the residue theorem more powerful than the generalized Cauchy integral formula? If so, how? In fact, if $f(z)$ has an Nth-order pole inside a piecewise smooth simple closed counterclockwise curve C, at $z = a$, then both the residue theorem and the generalized Cauchy integral formula give

$$I = \oint_C f(z)\,dz = 2\pi i \frac{1}{(N-1)!} \lim_{z\to a} \left\{ \frac{d^{N-1}}{dz^{N-1}} \left[(z-a)^N f(z) \right] \right\}. \tag{87}$$

However, it is traditional to rely on the residue theorem, for two reasons. First, the residue theorem applies to essential singularities as well as poles, whereas the generalized Cauchy integral formula applies only to poles. Second, and admittedly more superficial, is that the residue theorem is easily remembered: the integral is equal to $2\pi i$ times the sum of the residues.

EXERCISES 24.5

1. Let C_1 be a closed rectangular contour, traversed counterclockwise, with vertices at $-1-i, 3-i, 3+3i, -1+3i$. Let C_2 be a closed triangular contour, traversed clockwise, with vertices at -2, 2, and $-2+3i$. Evaluate the given integral by means of the residue theorem.

(a) $\displaystyle\oint_{C_1} \frac{dz}{\sin 2z}$ (b) $\displaystyle\oint_{C_1} \frac{dz}{z^2 e^z}$

(c) $\displaystyle\oint_{C_1} \frac{z^2 dz}{\sinh 2z}$ (d) $\displaystyle\oint_{C_1} \left(\frac{z+1}{z-1}\right)^3 dz$

(e) $\displaystyle\oint_{C_2} \frac{dz}{z^2-2iz-2}$ (f) $\displaystyle\oint_{C_2} \frac{dz}{\cosh^2(\pi z/2)}$

2. Evaluate by means of the residue theorem.

(a) $\displaystyle\int_0^\infty \frac{dx}{x^4+a^4}$ HINT: The zeros of $z^4 + a^4 = 0$ are $ae^{\pi i/4}$, $ae^{3\pi i/4}$, $ae^{5\pi i/4}$, and $ae^{7\pi i/4}$. Denote them as z_1, z_2, z_3, z_4, respectively. To work out the residue of $1/(z^4 + a^4)$ at z_1, for instance, it will be easier to evaluate

$$\lim_{z\to z_1} \left[(z-z_1)\frac{1}{z^4+a^4} \right]$$

by *l'Hôpital's rule* than to cancel the $(z - z_1)$'s and evaluate $1/[(z_1-z_2)(z_1-z_3)(z_1-z_4)]$. That step immediately gives the residue as $1/(4z_1^3)$.

(b) $\displaystyle\int_0^\infty \frac{dx}{(x^2+a^2)(x^2+b^2)}$ $(a > 0,\ b > 0)$

(c) $\displaystyle\int_0^\infty \frac{x^2}{x^4+1}\,dx$

(d) $\displaystyle\int_0^\infty \frac{dx}{(x^2+1)^2}$

(e) $\displaystyle\int_{-\infty}^{\infty} \frac{dx}{4x^2+2x+1}$

(f) $\displaystyle\int_{0}^{\infty} \frac{x^2}{x^6+1}\,dx$

(g) $\displaystyle\int_{0}^{\infty} \frac{\cos 2x}{(x^2+1)^2}\,dx$

(h) $\displaystyle\int_{0}^{\infty} \frac{x\sin x}{x^4+16}\,dx$

(i) $\displaystyle\int_{-\infty}^{\infty} \frac{\cos x}{8x^2+12x+5}\,dx$

(j) $\displaystyle\int_{-\infty}^{\infty} \frac{\sin 3x}{2x^2+2x+1}\,dx$

3. Evaluate the given integral by means of the residue theorem. HINT: The contour in the z plane must be a closed loop, one time around. Thus, the limits on the polar angle must be any 2π interval, such as 0 to 2π or $-\pi$ to π. Thus, in (a) for instance, use the evenness of $\sin^2 x$ to first rewrite the integral as $\frac{1}{2}\int_{-\pi}^{\pi}\sin^2 x\,dx$.

(a) $\displaystyle\int_{0}^{\pi} \sin^2 x\,dx$ (b) $\displaystyle\int_{0}^{\pi} \cos^2 x\,dx$

(c) $\displaystyle\int_{0}^{\pi/2} \sin^2 x\,dx$ (d) $\displaystyle\int_{\pi/2}^{\pi} \cos^2 x\,dx$

(e) $\displaystyle\int_{0}^{\pi} \sin^4 x\,dx$ (f) $\displaystyle\int_{0}^{\pi} \cos^4 x\,dx$

(g) $\displaystyle\int_{0}^{\pi} \sin^6 x\,dx$ (h) $\displaystyle\int_{0}^{4\pi} \cos^6 x\,dx$

(i) $\displaystyle\int_{-\pi}^{\pi} \frac{dt}{7+\cos t}$ (j) $\displaystyle\int_{0}^{\pi} \frac{dt}{1+\cos^2 t}$

(k) $\displaystyle\int_{0}^{2\pi} \frac{d\theta}{1+\sin^2\theta}$ (l) $\displaystyle\int_{0}^{2\pi} \frac{d\theta}{2+\sin 2\theta}$

4. Using the residue theorem, show that

$$\int_0^\pi \frac{\cos t\,dt}{1-2a\cos t+a^2} = \frac{\pi a}{1-a^2}. \qquad (-1<a<1)$$

5. Use the inversion formula and the residue theorem to evaluate the inverse of each of these Laplace transforms, where $a>0$ and $b>0$.

(a) $\dfrac{1}{s^2}$ (b) $\dfrac{1}{s^6}$ (c) $\dfrac{1}{s^2+a^2}$

(d) $\dfrac{1}{s^2-a^2}$ (e) $\dfrac{1}{(s-a)^4}$ (f) $\dfrac{s}{(s^2+a^2)^2}$

(g) $\dfrac{1}{(s-a)^2+b^2}$ (h) $\dfrac{e^{-as}}{s^3}$ (i) $\dfrac{e^{-as}}{(s-b)^2}$

6. Evaluate each by the residue theorem.

(a) $\displaystyle\int_{0}^{\infty} \frac{x^{a-1}}{x+1}\,dx \quad (0<a<1)$

(b) $\displaystyle\int_{0}^{\infty} \frac{\sqrt{x}}{x^3+1}\,dx$ (c) $\displaystyle\int_{0}^{\infty} \frac{dx}{\sqrt{x}(x^2+1)^2}$

(d) $\displaystyle\int_{0}^{\infty} \frac{\ln x}{x^2+1}\,dx$ (e) $\displaystyle\int_{0}^{\infty} \frac{x^{1/3}}{x^2+4}\,dx$

7. Evaluate

$$\int_0^\infty e^{-x^2}\cos 2ax\,dx = \frac{\sqrt{\pi}}{2}e^{-a^2} \qquad (a>0)$$

by integrating e^{-z^2} around a rectangle with vertices at 0, R, $R+ia$ and ia, and using the known integral

$$\int_0^\infty e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2}.$$

8. ($\int_0^\infty f(x)\,dx$ *where* $f(x)$ *is not even*) We saw in the examples that if $f(x)$ is even then $\int_0^\infty f(x)\,dx$ can be evaluated by first re-expressing it as $\frac{1}{2}\int_{-\infty}^{\infty} f(x)\,dx$. In this exercise we show what to do if $f(x)$ is not even. To illustrate, we will consider

$$I = \int_0^\infty \frac{dx}{x^2+x+1}.$$

(a) Show that considering

$$J = \oint_C \frac{dz}{z^2+z+1},$$

where C is the closed contour shown in Fig. 2 is of no help because using the residue theorem and letting $R\to\infty$ gives one equation in the *two* unknown integrals

$$\int_0^\infty \frac{dx}{x^2+x+1} \quad\text{and}\quad \int_0^\infty \frac{dx}{x^2-x+1}.$$

(b) Show that I can, however, be evaluated by inserting $\log z$ in the integrand and considering

$$J = \oint_C \frac{\log z}{z^2+z+1}\,dz,$$

where $\log z$ is defined by the branch cut in Fig. 4a and C is the contour shown in Fig. 4b; let $\epsilon\to 0$ and $R\to\infty$. Thus, show that $I = 2\pi/(3\sqrt{3})$.

9. Use the idea put forward in Exercise 8(b) to evaluate these integrals:

(a) $\displaystyle\int_0^\infty \frac{dx}{x^3+1}$ (b) $\displaystyle\int_0^\infty \frac{dx}{(x^2+2x+10)^2}$

(c) $\displaystyle\int_0^\infty \frac{x\,dx}{x^3+1}$ (d) $\displaystyle\int_0^1 \frac{dx}{x^2+1}$

HINT: For (d), set $t=(1-x)/x$.

10. (*Inversion of Fourier transforms*) Use the inversion formula and the residue theorem to evaluate the inverse of the given transform.

(a) $\dfrac{1}{\omega^2+i\omega+2}$ (b) $\dfrac{1}{\omega^2-3i\omega-2}$

(c) $\dfrac{1}{\omega^2+3i\omega-2}$ (d) $\dfrac{1}{(2-i\omega)^2}$

(e) $\dfrac{1}{(1+i\omega)^2}$ (f) $\dfrac{1}{(1+i\omega)^3}$

(g) $\dfrac{1}{(\omega^2+1)^2}$ (h) $\dfrac{1}{(\omega^2+4)^3}$

11. (*Cauchy principal value*) In Section 4.5 we defined the singular integral $\int_a^b f(x)\,dx$, where f is unbounded as $x\to a$, as the limit of the sequence of regular integrals $\int_{a+\epsilon}^b f(x)\,dx$ as $\epsilon\to 0$. The case where f is unbounded at an interior point was not considered there. In this exercise we consider the integral $\int_a^c f(x)\,dx$, where $|f(x)|\to\infty$ as $x\to b$, with $a<b<c$. In the same spirit as the limit definition given above, one defines

$$\int_a^c f(x)\,dx = \lim_{\substack{\epsilon_1\to 0\\ \epsilon_1\to 0}}\left[\int_a^{b-\epsilon_1} f(x)\,dx + \int_{b+\epsilon_2}^c f(x)\,dx\right], \tag{11.1}$$

where ϵ_1 and ϵ_2 tend to zero independently. If the limit exists, we say that $\int_a^c f(x)\,dx$ *exists*, or *converges*; otherwise, it *does not exist*, or *diverges*. Now, it is possible that the stated limit fails to exist, but does exist if we restrict the ϵ's to tend to zero together: that is, if $\epsilon_1=\epsilon_2$ ($=\epsilon$, say). In that case we write

$$\boxed{\oint_a^c f(x)\,dx = \lim_{\epsilon\to 0}\left[\int_a^{b-\epsilon} f(x)\,dx + \int_{b+\epsilon}^c f(x)\,dx\right],} \tag{11.2}$$

and call $\oint_a^c f(x)\,dx$ the **Cauchy principal value** of the integral. In place of the $\oint$ notation, some authors use $PV\int_a^c f(x)\,dx$.

(a) Show that

$$\int_{-1}^3 \frac{dx}{x}$$

is divergent [i.e., in the sense of (11.1)], but convergent in the Cauchy principal-value sense. Determine its Cauchy principal value.

(b) Repeat part (a) for $\int_1^4 dx/[x(x-2)]$.

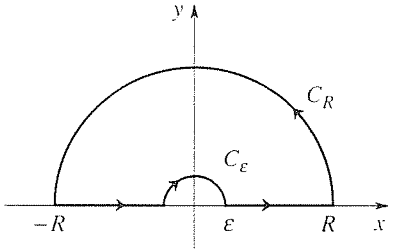
12. The integral

$$I=\int_0^\infty \frac{\sin x}{x}\,dx=\frac{\pi}{2} \tag{12.1}$$

is well known, and there are several ways of evaluating it. Here, we ask you to evaluate it using the residue theorem. HINT: Consider

$$J=\oint_C \frac{e^{iz}}{z}\,dz, \tag{12.2}$$

where C is the contour shown in the accompanying figure.



The circular indentation of radius ϵ is needed to avoid having the first-order pole of e^{iz}/z, at $z=0$, lie on the path of integration. (The function $\sin z/z$ is analytic everywhere but, as in Example 5, we consider e^{iz}/z instead so that the C_R integral will tend to zero as $R\to\infty$.) Then, show that

$$J=0=\oint_{-\infty}^{\infty}\frac{\cos x+i\sin x}{x}\,dx + \lim_{R\to\infty}\int_{C_R}\frac{e^{iz}}{z}\,dz + \lim_{\epsilon\to 0}\int_{C_\epsilon}\frac{e^{iz}}{z}\,dz, \tag{12.3}$$

where C_R and C_ϵ are the semicircular contours of radius R and ϵ, respectively. Show that the ML bound gives

$$\left|\int_{C_R}\frac{e^{iz}}{z}\,dz\right|\le\frac{1}{R}\pi R=\pi, \tag{12.4}$$

which is simply not sharp enough to show that $\int_{C_R}\to 0$ as $R\to\infty$. Thus, and this is the first time the ML bound has not sufficed, use the sharper bound given in Exercise 7 of

Section 24.2 to show that

$$\begin{aligned}\left|\int_{C_R} \frac{e^{iz}}{z}\,dz\right| &\le \int_0^\pi \left|\frac{e^{iz}}{z}\right| R\,d\theta \\ &= \int_0^\pi e^{-R\sin\theta}\,d\theta = 2\int_0^{\pi/2} e^{-R\sin\theta}\,d\theta \\ &< 2\int_0^{\pi/2} e^{-2R\theta/\pi}\,d\theta \\ &= \frac{\pi}{R}(1-e^{-R}) \to 0\end{aligned} \tag{12.5}$$

as $R \to \infty$, explaining each step. Turning to the C_ϵ integral, observe that

$$\begin{aligned}\frac{e^{iz}}{z} &= \frac{1}{z} + i - \frac{z}{2} + \cdots \qquad (0 < |z| < \infty) \\ &= \frac{1}{z} + g(z),\end{aligned} \tag{12.6}$$

and show that $g(z)$ is analytic for all z. Thus,

$$\begin{aligned}\lim_{\epsilon\to 0}\int_{C_\epsilon} \frac{e^{iz}}{z}\,dz &= \lim_{\epsilon\to 0}\left[\int_{C_\epsilon} \frac{1}{z}\,dz + \int_{C_\epsilon} g(z)\,dz\right] \\ &= -\pi i + \lim_{\epsilon\to 0}\int_{C_\epsilon} g(z)\,dz.\end{aligned} \tag{12.7}$$

Show that the latter limit is zero so that the limit of $\int_{C_\epsilon} e^{iz}\,dz/z$, as $\epsilon \to 0$, is $-\pi i$. Finally, note that

$$\fint_{-\infty}^{\infty} \frac{\cos x + i\sin x}{x}\,dx = \fint_{-\infty}^{\infty} \frac{\cos x}{x}\,dx + i\int_{-\infty}^{\infty} \frac{\sin x}{x}\,dx, \tag{12.8}$$

where the latter integral does not need the Cauchy principal-value sign because the integral is not singular; that is, whereas $(\cos x)/x \sim 1/x$ as $x \to 0$, $(\sin x)/x \sim 1$ there.

Chapter 24 Review

After introducing the concept of complex series in Section 24.2.1, we use the Cauchy integral formula to derive Taylor series in Section 24.2.2. Whereas the Taylor series of $f(x)$ about $x = a$ converges in an *interval* $|x - a| < R$ (or perhaps only at the point $x = a$), the Taylor series of $f(z)$ about $z = a$ converges in a *disk* $|z - a| < R$ (or perhaps only at the point $z = a$), the radius of which is the distance from a to the nearest singularity of f. Theorem 24.2.8 reveal an intimate connection between Taylor series and power series: if a power series converges in $|z - a| < R$, then its sum function $f(z)$ is analytic there and, indeed, the power series is the Taylor series of $f(z)$.

Besides the Taylor series representation of f in the disk $|z - a| < R$, one can develop power series, called Laurent series, that represent f in *annuli* of analyticity. Laurent series differ from Taylor series in that they necessarily include one or more negative powers of $z - a$, and in that their coefficients are given by the unwieldy integral expression $c_n = (1/2\pi i)\oint_C f(\zeta)\,d\zeta/(\zeta - a)^{n+1}$. However, we showed that we can bypass that formula and determine the c_n's using only the methods of Taylor series.

With Laurent series in hand, we are then able to make precise the notion of singularities, by categorizing them into different types. Distinguishing, first, isolated singularities from nonisolated ones, we limit our attention to the former, which are of chief interest to us. Among isolated singularities, we distinguish Nth-order poles and essential singularities.

Finally, we are able to return to contour integration and use Laurent series to derive the powerful and elegant residue theorem which states, essentially, that if $f(z)$ has only isolated singularities within C, then $\oint_C f(z)\,dz$ is simply $2\pi i$ times the sum of the residues within C, where the residue of f at a singular point z_j is the c_{-1} coefficient in the Laurent series of f about z_j, within an annulus $0 < |z - z_j| < \rho$. Besides its obvious applicability to integrals on closed paths in the z plane, the residue theorem can, as we see in the examples and exercises, be applied to real integrals and Laplace and Fourier inversions.

References

The following references are recommended for collateral reading and for further study. The list is by no means complete.

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. NY: Dover, 1965. Comprehensive collection of formulas, properties, and tables of values of the special functions of mathematical physics.

2. T. M. Apostol, *Mathematical Analysis*. Reading, MA: Addison–Wesley, 1957. An excellent general and rigorous reference for advanced calculus.

3. C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*. NY: McGraw–Hill, 1978. Covers more advanced topics, including methods of solution of differential equations with irregular singular points.

4. R. V. Churchill, *Operational Mathematics*, 3rd ed. NY: McGraw–Hill, 1972. A standard reference for the theory and application of the Laplace transform.

5. I. Ekeland, *Mathematics and the Unexpected*. Chicago: University of Chicago Press, 1988. A stimulating historical and mathematical account of the profound shift from the orderly determinism of Newton to an often nondeterministic and chaotic world. This little book forms an excellent sequel to the Kline reference listed below.

6. A. Erdèlyi (ed), *Tables of Integral Transforms*, Vol. 1. NY: McGraw–Hill, 1954. Volume 1 contains a tabulation of Fourier transforms (called "exponential Fourier transforms" therein), Fourier sine and cosine transforms, Laplace transforms, and Mellin transforms.

7. M. D. Greenberg, *Foundations of Applied Mathematics*. Upper Saddle River, NJ: Prentice Hall, 1978. Chapter 10 on the variational calculus discusses the derivation of ordinary and partial differential equations from fundamental extremum principles such as the physical principle of minimum potential energy. Chapter 25 gives an introduction to perturbation methods, Chapter

27 explains the method of characteristics for the numerical solution of hyperbolic PDE's, and Chapter 28 explains the method of Green's functions for solving ODE's and PDE's.

8. E. L. Ince, *Ordinary Differential Equations*. NY: Dover, 1956. Since its first publication in 1926 this has been a classic reference work.

9. D. E. Johnson and J. R. Johnson, *Mathematical Methods in Engineering Physics*. Upper Saddle River, NJ: Prentice Hall, 1982. Includes a readable account of the numerous special functions of mathematical physics.

10. D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, 2nd ed. Oxford: Oxford University Press, 1987.

11. M. Kline, *Mathematical Thought from Ancient to Modern Times*. NY: Oxford University Press, 1972. An interesting historical account of the roots of modern mathematics.

12. N. W. McLachlan, *Bessel Functions for Engineers*, 2nd ed. London: Oxford University Press, 1955. Indicates the wide variety of engineering applications of Bessel functions.

13. G. M. Murphy, *Ordinary Differential Equations and Their Solutions*. Princeton, NJ: Van Nostrand, 1960. Contains an extensive compilation of ordinary differential equations and their solutions.

14. H. E. Newell, Jr., *Vector Analysis*. NY: McGraw-Hill, 1955. Chapter 11 contains a 43-page introduction to the application of vector analysis to electromagnetic theory.

15. E. G. Phillips, *Functions of a Complex Variable*. Edinburgh: Oliver & Boyd, 1957. An excellent little book with many worked examples, especially on contour integration.

16. D. L. Powers, *Boundary Value Problems*. NY: Academic Press, 1972. A readable text for a first course in PDE's.

17. E. B. Saff and A. D. Snider, *Fundamentals of Complex Analysis*. Upper Saddle River, NJ: Prentice Hall, 1976.

18. H. Sagan, *Boundary and Eigenvalue Problems in Mathematical Physics*. NY: John Wiley, 1961. Detailed treatment of Sturm–Liouville and differential equation eigenvalue problems.

19. H. M. Schey, *Div, Grad, Curl, and All That*. NY: Norton, 1973.

20. W. G. Strang, *Linear Algebra and Its Applications*, 2nd ed. NY: Academic Press, 1980. Clear, with engaging applications and an introduction to linear programming.

21. S. H. Strogatz, *Nonlinear Dynamics and Chaos*. Reading, MA: Addison–Wesley, 1994. A more detailed development of the material presented here in Chapter 7.

22. E. Zauderer, *Partial Differential Equations of Applied Mathematics*. NY: John Wiley, 1983. This graduate-level text includes nonlinear problems, and perturbation and asymptotic methods, as well as discussion of the method of characteristics for the numerical solution of hyperbolic PDE's.

Appendix A

Review of Partial Fraction Expansions

Generally, one meets the method of partial fraction expansions in the integral calculus, where the method is used to express a difficult integral as a linear combination of simpler ones. For example,

$$\begin{aligned}\int \frac{dx}{x^2+3x+2} &= \int \left(\frac{1}{2}\frac{1}{x+1} - \frac{1}{2}\frac{1}{x+3}\right) dx \\ &= \frac{1}{2}\int \frac{dx}{x+1} - \frac{1}{2}\int \frac{dx}{x+3} \\ &= \frac{1}{2}\ln|x+1| - \frac{1}{2}\ln|x+3| + \text{constant}.\end{aligned}$$

In this text we use the method primarily to help us to invert Laplace and Fourier transforms such as the Laplace transform $F(s) = 1/(s^2+3s+2)$. For convenient reference, this appendix contains a review of the method.

Let $p(x)$ and $q(x)$ be finite-degree polynomials in x, of degree P and Q, respectively. Then

$$f(x) = \frac{p(x)}{q(x)} \tag{A1}$$

is called a **rational function** of x. Let P be less than Q. [If $P \geq Q$, then we can, by the long division of q into p, express f as a polynomial of degree $P - Q$ plus a rational function $r(x)/q(x)$, where the degree R of r *is* less than Q. For instance, long division gives

$$\frac{x^5+6x^2-5x+6}{x^3-x^2+x+1} = x^2 - x + 2 + \frac{x^2-4x+1}{x^3+x^2-x+3}.$$

Whereas the method of partial fractions cannot be applied to the rational function on the left (because $P = 5$ is not less than $Q = 3$), it can be applied to the one on the right (because $P = 2$ *is* less than $Q = 3$).]

Distinct roots. Let $q(x)$ in (A1) have the distinct roots $x_1, \ldots, x_Q$. Then f admits the **partial fraction expansion**

$$\boxed{f(x) = \frac{p(x)}{q(x)} = \frac{a_1}{x - x_1} + \frac{a_2}{x - x_2} + \cdots + \frac{a_Q}{x - x_Q},} \tag{A2}$$

where the a_j's are constants. One way to determine the a_j's is to recombine the terms on the right-hand side over a common denominator [namely, $q(x)$] and require its numerator to be identical to $p(x)$.

EXAMPLE 1. Expand $f(x) = (x - 1)/(x^2 + 5x + 6)$ in partial fractions. Since $x^2 + 5x + 6 = (x + 2)(x + 3)$, we can expand f as

$$f(x) = \frac{x - 1}{x^2 + 5x + 6} = \frac{a_1}{x + 2} + \frac{a_2}{x + 3}.$$

To determine a_1 and a_2, write

$$\frac{x - 1}{x^2 + 5x + 6} = \frac{(a_1 + a_2)x + (3a_1 + 2a_2)}{x^2 + 5x + 6}.$$

For the numerators to be identical we need

$$\begin{aligned} x^0 &: \quad -1 = 3a_1 + 2a_2, \\ x^1 &: \quad\ \ 1 = a_1 + a_2. \end{aligned}$$

Solving these equations gives $a_1 = -3$ and $a_2 = 4$ so

$$\frac{x - 1}{x^2 + 5x + 6} = -\frac{3}{x + 2} + \frac{4}{x + 3}. \quad ▮$$

However, it is simpler to proceed as follows. To calculate a_1, multiply (A2) by $x - x_1$, then let $x \to x_1$ in the result. That step gives

$$\lim_{x \to x_1} \left[(x - x_1) \frac{p(x)}{q(x)} \right] = a_1 + 0 + \cdots + 0.$$

To find a_2 multiply by $x - x_2$ and let $x \to x_2$, and so on. Thus,

$$a_j = \lim_{x \to x_j} \left[(x - x_j) \frac{p(x)}{q(x)} \right] \tag{A3}$$

or, applying l'Hôpital's rule to the indeterminate part, $(x - x_j)/q(x)$,

$$\boxed{a_j = \frac{p(x_j)}{q'(x_j)}.} \tag{A4}$$

In Example 1, $x_1 = -2$ and $x_2 = -3$ so (A4) gives $a_1 = [(x-1)/(2x+5)]|_{x=-2} = -3$ and $a_2 = [(x-1)/(2x+5)]|_{x=-3} = 4$ as obtained above. In summary, if $q(x)$ has distinct roots, then $f(x) = p(x)/q(x)$ can be expanded in partial fractions according to (A2) and the a_j's can be found readily according to (A4).

Repeated roots. If any root x_j of $q(x)$ is of multiplicity k [i.e., $(x-x_j)^k$ is a factor of $q(x)$], then the jth term on the right-hand side of (A2) must be modified to the form

$$\boxed{\frac{a_{j1}}{x-x_j} + \frac{a_{j2}}{(x-x_j)^2} + \cdots + \frac{a_{jk}}{(x-x_j)^k}} \tag{A5}$$

or, equivalently,

$$\frac{b_{j0} + b_{j1}x + \cdots + b_{j,k-1}x^{k-1}}{(x-x_j)^{k-1}}.$$

To solve for $a_{j1}, \dots, a_{jk}$ in (A5), we can recombine terms over a common denominator [namely, $q(x)$] and equate coefficients of powers of x in the numerator (since powers of x are linearly independent functions of x) as we did in Example 1.

EXAMPLE 2. To expand $(4x^2+5)/[(x-2)^3(x+3)]$ in partial fractions, write

$$\begin{aligned}\frac{4x^2+5}{(x-2)^3(x+3)} &= \left[\frac{a}{x-2} + \frac{b}{(x-2)^2} + \frac{c}{(x-2)^3}\right] + \frac{d}{x+3}\\ &= \frac{\left[a(x-2)^2 + b(x-2) + c\right](x+3) + d(x-2)^3}{(x-2)^3(x+3)}\\ &= [(a+d)x^3 + (-a+b-6d)x^2 + (-8a+b+c+12d)x\\ &\quad + (12a-6b+3c-8d)]/[(x-2)^3(x+3)],\end{aligned}$$

where the notation a, b, c, d will be simpler than using subscripted a_{jk}'s. Thus,

$$\begin{aligned} x^0: &\quad 5 = 12a - 6b + 3c - 8d,\\ x^1: &\quad 0 = -8a + b + c + 12d,\\ x^2: &\quad 4 = -a + b - 6d,\\ x^3: &\quad 0 = a + d,\end{aligned}$$

with solution $a = 41/125$, $b = 59/25$, $c = 21/5$, $d = -41/125$ so

$$\frac{4x^2+5}{(x-2)^3(x+3)} = \frac{41}{125}\frac{1}{x-2} + \frac{59}{25}\frac{1}{(x-2)^2} + \frac{21}{5}\frac{1}{(x-2)^3} - \frac{41}{125}\frac{1}{x+3}. \quad ▮$$

In Example 2 we could have used (A4) to compute d but we could not have used it to compute a, b, c because $x = 2$ is a repeated root. To compute $a_{j1}, \dots, a_{jk}$ in (A5), we need a modified version of (A4), namely,

$$\boxed{a_{jm} = \frac{1}{(k-m)!}\frac{d^{(k-m)}}{dx^{(k-m)}}\left[(x-x_j)^k\frac{p(x)}{q(x)}\right]\Bigg|_{x\to x_j}} \tag{A6}$$

for $m = 1, \ldots, k$. For instance, if we apply (A6) to the calculation of a, b, c in Example 2 we obtain

$$a = a_{11} = \frac{1}{2!}\frac{d^2}{dx^2}\left[(x-2)^3\frac{4x^2+5}{(x-2)^3(x+3)}\right]\bigg|_{x\to 2} = \frac{41}{125},$$

$$b = a_{12} = \frac{1}{1!}\frac{d}{dx}\left[(x-2)^3\frac{4x^2+5}{(x-2)^3(x+3)}\right]\bigg|_{x\to 2} = \frac{59}{25},$$

and

$$c = a_{13} = \frac{1}{0!}\left[(x-2)^3\frac{4x^2+5}{(x-2)^3(x+3)}\right]\bigg|_{x\to 2} = \frac{21}{5},$$

which results agree with those obtained above.

Partial fraction expansions can also be carried out using computer software. For example, the relevant *Maple* command is **convert**, and the commands

$$\text{convert}((x-1)/(x\hat{\ }2+5*x+6), \text{parfrac}, x);$$

and

$$\text{convert}((4*x\hat{\ }2+5)/((x-2)\hat{\ }3*(x+3)), \text{parfrac}, x);$$

give the results that we obtained in Examples 1 and 2, respectively.

We leave the derivation of (A6) for you as an exercise.

Appendix B

Existence and Uniqueness of Solutions of Systems of Linear Algebraic Equations

This appendix is intended as a minimal prerequisite for Chapter 3. Alternatively, to integrate linear algebra more heavily with the ODE chapters, we suggest following Chapter 2 with Sections 8.1–10.6 before beginning Chapter 3.

Definitions. We call the m equations

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= c_1, \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= c_2, \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= c_m,
\end{aligned} \tag{B1}$$

where the a_{jk} coefficients and the c_j's are known, a **system** of m linear algebraic equations on the n unknowns $x_1, \ldots, x_n$. Any set of numbers $x_1, \ldots, x_n$ that renders each of the m equations a numerical equality is called a **solution** of the system. A system is said to be **consistent** if it admits at least one solution, and **inconsistent** if it admits none. It is shown in Chapter 8 that if (B1) is consistent, then it admits either a **unique** solution (one solution) or an infinity of them, never three solutions or 27 solutions, for instance.

We call the array of coefficients

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{B2}$$

the **coefficient matrix**, and enclose the a_{jk} elements between brackets simply to show that the entire array is being regarded as a single entity. We say that **A** has

m rows and n columns. For instance, the second row consists of the elements $a_{21}, a_{22}, \ldots, a_{2n}$.

The case where $m{=}n$. In applications, the number of equations (m) is usually, but not always, equal to the number of unknowns (n); in general, m can be less than, equal to, or greater than n. Consider first the case where $m = n$. As the simplest case, let $m = n = 1$ so (B1) becomes $a_{11}x_1 = c_1$, or simply

$$ax = c. \tag{B3}$$

If $a \neq 0$, then (B3) admits a unique solution, namely, $x = c/a$. If $a = 0$, however, then there are two possibilities, as can be seen from (B3): if $c \neq 0$, then there is no solution and (B3) is inconsistent, and if $c = 0$, then (B3) is consistent and there is the infinity of solutions $x = \alpha$, where α is arbitrary.

The upshot, for this simple case, is that whether or not the coefficient a is zero is crucial: if $a \neq 0$ (the generic case), then there is a unique solution, and if $a = 0$ (the nongeneric or *singular* case) then there is either no solution or an infinity of them, depending upon the value of c. This idea generalizes to the case where $m = n > 1$ as indicated in (i) and (ii) below. First, we define a so-called **determinant** of the **A** matrix, and denote it as

$$\det\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}, \tag{B4}$$

that is, with straight line braces instead of square brackets. The determinant of **A** is a number (which can be positive, negative, or zero), defined as

$$\begin{vmatrix} a_{11} \end{vmatrix} \equiv a_{11}, \tag{B5a}$$

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \equiv a_{11}a_{22} - a_{21}a_{12}, \tag{B5b}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \equiv \begin{aligned} & a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{21}a_{33} - a_{31}a_{23}) \\ & + a_{13}(a_{21}a_{32} - a_{31}a_{22}) \end{aligned} \tag{B5c}$$

for $n = 1, 2$, and 3, respectively. The general definition for any n is given in Section 10.4, but the cases (B5a,b,c) should suffice for Chapter 3. Do not confuse determinant with absolute value, especially in (B5a) where we cannot tell, from the left-hand side, whether we are signifying the absolute value of a_{11} or the determinant of the tiny matrix $\mathbf{A}= [a_{11}]$, without being told or from the context. For instance, $|-6| = -6$, $\begin{vmatrix} 2 & 5 \\ -3 & 4 \end{vmatrix} = 8 - (-15) = 23$, and

$$\begin{vmatrix} 3 & -2 & 5 \\ 0 & 4 & 6 \\ 1 & 1 & 8 \end{vmatrix} = (3)(32-6)-(-2)(0-6)+(5)(0-4)=46.$$

Now, whether or not det**A** is zero is crucial insofar as the existence and uniqueness of solutions of the system (B1):

(i) If det**A** $\neq 0$, then (B1) is consistent and has a unique solution.

(ii) If det**A** $= 0$, then (B1) has either no solution (is inconsistent) or an infinity of solutions, depending upon the c_j's.

For instance, consider the three systems

$$\begin{aligned} 2x_1 - x_2 &= 4 \\ 3x_1 + x_2 &= 11, \end{aligned} \qquad \begin{aligned} 2x_1 - x_2 &= 5 \\ 4x_1 - 2x_2 &= 3, \end{aligned} \quad \text{and} \quad \begin{aligned} 2x_1 - x_2 &= 5 \\ 4x_1 - 2x_2 &= 10. \end{aligned}$$

In the first case, det**A** $= 5 \neq 0$ so there is a unique solution (namely, $x_1 = 3$ and $x_2 = 2$); in the second case, det**A** $= 0$ and there is no solution (as is not surprising since the second left-hand side is twice the first left-hand side whereas 3 is not twice 5); and in the third case, det**A** $= 0$ and there is an infinity of solutions [namely, $x_2 = \alpha$ and $x_1 = (5+\alpha)/2$, where α is arbitrary].

In fact, if det**A** $\neq 0$, then we can give a formula for the solution of (B1). Namely, for each j from 1 to n, x_j is given as the ratio of two determinants: the one in the denominator is det**A**, and the one in the numerator is the same except that its jth column is replaced by the column of c's on the right-hand side of (B1). This result is known as **Cramer's rule**. For instance, if

$$\begin{aligned} 3x_1 - x_2 + x_3 &= 1 \\ 2x_1 - 2x_2 + x_3 &= 0 \\ 4x_1 + 3x_2 - x_3 &= -5, \end{aligned} \tag{B6}$$

then Cramer's rule gives

$$x_1 = \frac{\begin{vmatrix} 1 & -1 & 1 \\ 0 & -2 & 1 \\ -5 & 3 & -1 \end{vmatrix}}{\begin{vmatrix} 3 & -1 & 1 \\ 2 & -2 & 1 \\ 4 & 3 & -1 \end{vmatrix}} = \frac{-6}{5} = -\frac{6}{5}, \qquad x_2 = \frac{\begin{vmatrix} 3 & 1 & 1 \\ 2 & 0 & 1 \\ 4 & -5 & -1 \end{vmatrix}}{\begin{vmatrix} 3 & -1 & 1 \\ 2 & -2 & 1 \\ 4 & 3 & -1 \end{vmatrix}} = \frac{11}{5}, \tag{B7}$$

and, in similar fashion, $x_3 = 34/5$. Notice from Cramer's rule that since each x_j is given as the ratio of two determinants, with the determinant in each denominator being the determinant of **A**, it follows that if det**A** $\neq 0$, then there exists a unique solution for each x_j. That result is identical to (i) stated above.

Remember that (i) and (ii), and Cramer's rule, hold only in the case where $m = n$. If $m \neq n$, then the "determinant of **A**" is not even defined. More about this in Section 10.4.

The homogeneous case. If all of the c_j's in (B1) are zero, then we say that the system is homogeneous. It should be evident that *homogeneous systems are always consistent* since they necessarily admit the solution $x_1 = x_2 = \cdots = x_n = 0$, which is known as the **trivial solution**. (When we give it that name, we do not mean to imply that it is in some way beneath our dignity; it is a perfectly legitimate solution.) Further, *if (B1) is a homogeneous system with $m < n$, then it necessarily admits not only the trivial solution, but also an infinity of nontrivial solutions.* For instance, it can be verified by substitution that the system

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 0 \\ 2x_1 + x_2 - x_3 &= 0 \end{aligned} \tag{B8}$$

admits the solution $x_1 = x_3 = \alpha$ and $x_2 = -\alpha$ for any α (which solutions include the trivial solution for the choice $\alpha = 0$).

Appendix C

Table of Laplace Transforms

$f(t)$	$\overline{f}(s) = \int_0^\infty f(t)e^{-st}\,dt$
NOTE: s is regarded as real here.	
1. 1	$\frac{1}{s}$ $(s > 0)$
2. e^{at}	$\frac{1}{s-a}$ $(s > a)$
3. $\sin at$	$\frac{a}{s^2+a^2}$ $(s > 0)$
4. $\cos at$	$\frac{s}{s^2+a^2}$ $(s > 0)$
5. $\sinh at$	$\frac{a}{s^2-a^2}$ $(s > \lvert a \rvert)$
6. $\cosh at$	$\frac{s}{s^2-a^2}$ $(s > \lvert a \rvert)$
7. t^n $(n =$ positive integer$)$	$\frac{n!}{s^{n+1}}$ $(s > 0)$
8. t^p $(p > -1)$	$\frac{\Gamma(p+1)}{s^{p+1}}$ $(s > 0)$
9. $e^{at}\sin bt$	$\frac{b}{(s-a)^2+b^2}$ $(s > a)$
10. $e^{at}\cos bt$	$\frac{s-a}{(s-a)^2+b^2}$ $(s > a)$
11. $t\sin at$	$\frac{2as}{(s^2+a^2)^2}$ $(s > 0)$
12. $t\cos at$	$\frac{s^2-a^2}{(s^2+a^2)^2}$ $(s > 0)$
13. $t\sinh at$	$\frac{2as}{(s^2-a^2)^2}$ $(s > a)$

	$f(t)$	$\bar{f}(s)=\int_0^\infty f(t)e^{-st}\,dt$
14.	$t\cosh at$	$\dfrac{s^2+a^2}{(s^2-a^2)^2}\quad(s>a)$
15.	$t^n e^{at}\quad(n=\text{positive integer})$	$\dfrac{n!}{(s-a)^{n+1}}\quad(s>a)$
16.	$t^p e^{at}\quad(p>-1)$	$\dfrac{\Gamma(p+1)}{(s-a)^{p+1}}\quad(s>a)$
17.	$\ln t$	$-\dfrac{\gamma+\ln s}{s}\quad(s>0)$ $(\gamma=\text{Euler's constant}\approx 0.577215665)$
18.	$H(t-a)\quad(a\geq 0)$	$\dfrac{e^{-as}}{s}\quad(s>0)$
19.	$\delta(t-a)\quad(a>0)$	e^{-as}
20.	$\dfrac{e^{-a^2/t}}{\sqrt{t}}\quad(a\geq 0)$	$\sqrt{\dfrac{\pi}{s}}\,e^{-2a\sqrt{s}}\quad(s>0)$
21.	$\dfrac{e^{-a^2/t}}{t^{3/2}}\quad(a>0)$	$\dfrac{\sqrt{\pi}}{a}\,e^{-2a\sqrt{s}}\quad(s>0)$
22.	$J_0(t)$	$\dfrac{1}{\sqrt{s^2+1}}\quad(s>0)$
Linearity of transform and inverse (Theorems 5.3.1, 5.3.2):		
23.	$\alpha u(t)+\beta v(t)$	$\alpha\bar{u}(s)+\beta\bar{v}(s)$
Transform of derivative (Theorem 5.3.3):		
24.	$f'(t)$	$s\bar{f}(s)-f(0)$
25.	$f''(t)$	$s^2\bar{f}(s)-sf(0)-f'(0)$
26.	$f^{(n)}(t)$	$s^n\bar{f}(s)-s^{n-1}f(0)-s^{n-2}f'(0)-\cdots$ $-sf^{(n-2)}(0)-f^{(n-1)}(0)$
Transform of integral (Theorem 5.7.3):		
27.	$\int_0^t f(\tau)\,d\tau$	$\dfrac{\bar{f}(s)}{s}$
Laplace convolution theorem (Theorem 5.3.4):		
28.	$(f*g)(t)=\int_0^t f(\tau)g(t-\tau)d\tau$	$\bar{f}(s)\bar{g}(s)$

	$f(t)$	$\overline{f}(s) = \int_0^\infty f(t)e^{-st}\,dt$
s-Shift (Theorem 5.7.1):		
29.	$e^{-at}f(t)$	$\overline{f}(s+a)$
t-Shift (Theorem 5.7.2):		
30.	$H(t-a)f(t-a)$	$e^{-as}\overline{f}(s)$
Multiplication by t and $1/t$ (Theorems 5.7.4 and 5.7.5):		
31.	$tf(t)$	$-\dfrac{d\overline{f}(s)}{ds}$
32.	$\dfrac{f(t)}{t}$	$\int_s^\infty \overline{f}(s')\,ds'$
Transform of periodic function (Theorem 5.7.8):		
33.	$f(t)$, of period T	$\dfrac{1}{1-e^{-sT}}\int_0^T f(t)e^{-st}\,dt$

Appendix D

Table of Fourier Transforms

	$f(x)$	$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx$
1.	$\dfrac{1}{x^2 + a^2} \quad (a > 0)$	$\dfrac{\pi}{a} e^{-a\lvert\omega\rvert}$
2.	$H(x) e^{-ax} \quad (\operatorname{Re} a > 0)$	$\dfrac{1}{a + i\omega}$
3.	$H(-x) e^{ax} \quad (\operatorname{Re} a > 0)$	$\dfrac{1}{a - i\omega}$
4.	$e^{-a\lvert x\rvert} \quad (a > 0)$	$\dfrac{2a}{\omega^2 + a^2}$
5.	e^{-x^2}	$\sqrt{\pi} e^{-\omega^2/4}$
6.	$\dfrac{1}{2a\sqrt{\pi}} e^{-x^2/(2a)^2} \quad (a > 0)$	$e^{-a^2\omega^2}$
7.	$\dfrac{1}{\sqrt{\lvert x\rvert}}$	$\sqrt{\dfrac{2\pi}{\lvert\omega\rvert}}$
8.	$e^{-a\lvert x\rvert/\sqrt{2}} \sin\left(\dfrac{a}{\sqrt{2}}\lvert x\rvert + \dfrac{\pi}{4}\right) \quad (a > 0)$	$\dfrac{2a^3}{\omega^4 + a^4}$
9.	$H(x + a) - H(x - a)$	$\dfrac{2 \sin \omega a}{\omega}$
10.	$\delta(x - a)$	$e^{-i\omega a}$
11.	$f(ax + b) \quad (a > 0)$	$\dfrac{1}{a} e^{ib\omega/a} \hat{f}\left(\dfrac{\omega}{a}\right)$
12.	$\dfrac{1}{a} e^{-ibx/a} f\left(\dfrac{x}{a}\right) \quad (a > 0,\ b \text{ real})$	$\hat{f}(a\omega + b)$
13.	$f(ax) \cos cx \quad (a > 0,\ c \text{ real})$	$\dfrac{1}{2a}\left[\hat{f}\left(\dfrac{\omega - c}{a}\right) + \hat{f}\left(\dfrac{\omega + c}{a}\right)\right]$
14.	$f(ax) \sin cx \quad (a > 0,\ c \text{ real})$	$\dfrac{1}{2ai}\left[\hat{f}\left(\dfrac{\omega - c}{a}\right) - \hat{f}\left(\dfrac{\omega + c}{a}\right)\right]$
15.	$f(x + c) + f(x - c) \quad (c \text{ real})$	$2\hat{f}(\omega) \cos \omega c$

$f(x)$	$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx$
16. $f(x+c) - f(x-c) \quad (c\,\text{real})$	$2i\hat{f}(\omega)\sin \omega c$
17. $x^n f(x) \quad (n = 1, 2, \ldots)$	$i^n \dfrac{d^n}{d\omega^n}\, \hat{f}(\omega)$
Linearity of transform and inverse:	
18. $\alpha f(x) + \beta g(x)$	$\alpha \hat{f}(\omega) + \beta \hat{g}(\omega)$
Transform of derivative:	
19. $f^{(n)}(x)$	$(i\omega)^n \hat{f}(\omega)$
Transform of integral:	
20. $f(x) = \int_{-\infty}^{x} g(\xi)\, d\xi,$ where $f(x) \to 0$ as $x \to \infty$	$\hat{f}(\omega) = \dfrac{1}{i\omega}\, \hat{g}(\omega)$
Fourier convolution theorem:	
21. $(f * g)(x) = \int_{-\infty}^{\infty} f(x-\xi) g(\xi)\, d\xi$	$\hat{f}(\omega)\hat{g}(\omega)$

Appendix E

Table of Fourier Cosine and Sine Transforms

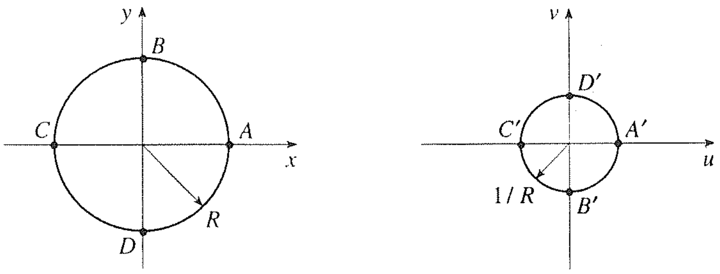
	$f(x)$	$\hat{f}_C(\omega) = \int_0^\infty f(x)\cos\omega x\,dx$
1C.	$e^{-ax} \quad (a > 0)$	$\frac{a}{\omega^2 + a^2}$
2C.	$x^n e^{-ax} \quad (a > 0)$	$\frac{n!\,\mathrm{Re}\,(a + i\omega)^{n+1}}{(\omega^2 + a^2)^{n+1}}$ (Re = real part)
3C.	$\frac{1}{x^2 + a^2} \quad (a > 0)$	$\frac{\pi}{2a} e^{-a\omega}$
Linearity of transform and inverse:		
4C.	$\alpha f(x) + \beta g(x)$	$\alpha \hat{f}_C(\omega) + \beta \hat{g}_C(\omega)$
Transform of derivative:		
5C.	$f'(x)$	$\omega \hat{f}_S(\omega) - f(0)$
6C.	$f''(x)$	$-\omega^2 \hat{f}_C(\omega) - f'(0)$
Convolution theorem:		
7C.	$\frac{1}{2}\int_0^\infty [f(\lvert x - \xi \rvert) + f(x + \xi)] g(\xi)\,d\xi$	$\hat{f}_C(\omega)\hat{g}_C(\omega)$

	$f(x)$	$\hat{f}_S(\omega) = \int_0^\infty f(x) \sin \omega x \, dx$
1S.	$e^{-ax} \quad (a > 0)$	$\dfrac{\omega}{\omega^2 + a^2}$
2S.	$x^n e^{-ax} \quad (a > 0)$	$\dfrac{n! \, \mathrm{Im} \, (a + i\omega)^{n+1}}{(\omega^2 + a^2)^{n+1}}$ (Im = imaginary part)
3S.	$\dfrac{x}{x^2 + a^2} \quad (a > 0)$	$\dfrac{\pi}{2} e^{-a\omega}$
Linearity of transform and inverse:		
4S.	$\alpha f(x) + \beta g(x)$	$\alpha \hat{f}_S(\omega) + \beta \hat{g}_S(\omega)$
Transform of derivative:		
5S.	$f'(x)$	$-\omega \hat{f}_C(\omega)$
6S.	$f''(x)$	$-\omega^2 \hat{f}_S(\omega) + \omega f(0)$
Convolution theorem:		
7S.	$\dfrac{1}{2} \int_0^\infty [f(\lvert x - \xi \rvert) - f(x + \xi)] g(\xi) \, d\xi$	$\hat{f}_C(\omega) \hat{g}_S(\omega)$

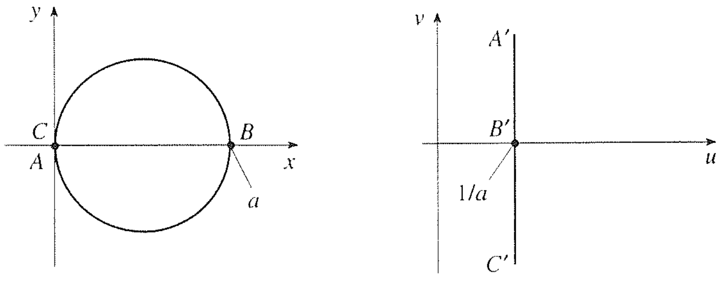
Appendix F

Table of Conformal Maps

1. $w = 1/z$

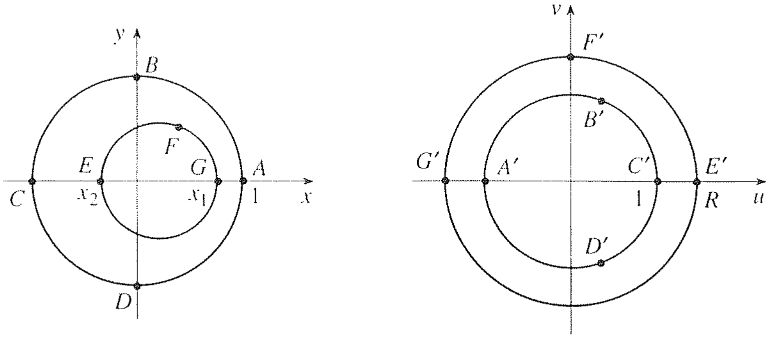


2. $w = 1/z$



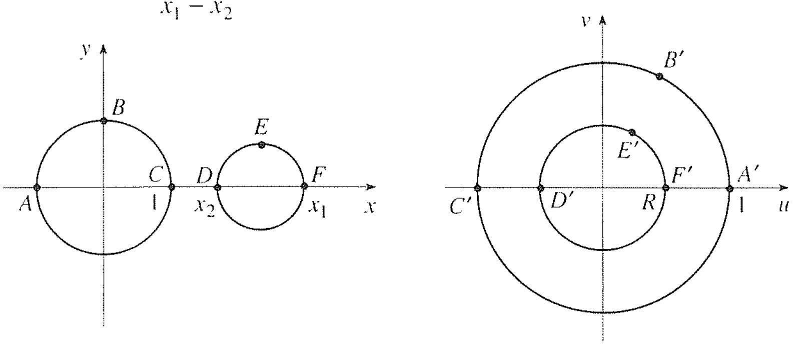
3. $w = \dfrac{z-a}{az-1}$; $a = \dfrac{1 + x_1x_2 + \sqrt{(1-x_1^2)(1-x_2^2)}}{x_1+x_2}$,

$$R = \frac{1 - x_1x_2 + \sqrt{(1-x_1^2)(1-x_2^2)}}{x_1 - x_2}, \quad -1 < x_2 < x_1 < 1$$

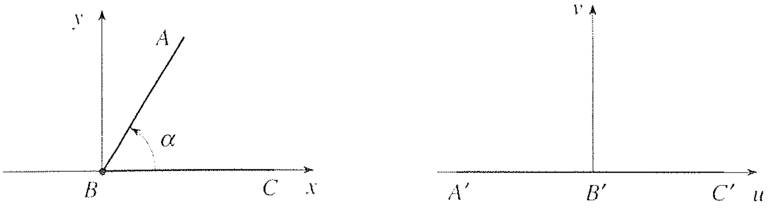


4. $w = \dfrac{z-a}{az-1}$; $a = \dfrac{x_1x_2 + 1 + \sqrt{(x_1^2-1)(x_2^2-1)}}{x_1+x_2}$,

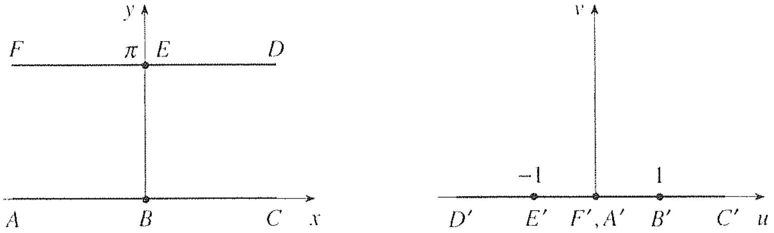
$$R = \frac{x_1x_2 - 1 - \sqrt{(x_1^2-1)(x_2^2-1)}}{x_1 - x_2}, \quad 1 < x_2 < x_1$$



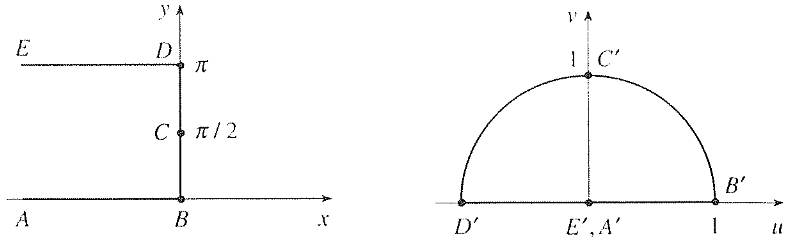
5. $w = z^{\pi/\alpha}$



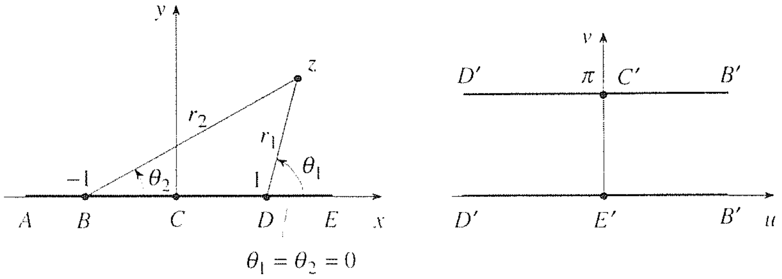
6. $w = e^z$



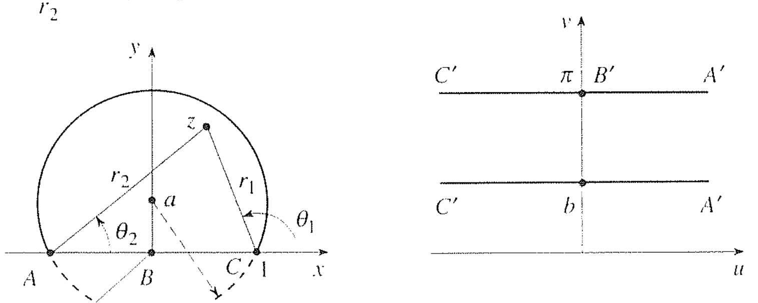
7. $w = e^z$



8. $w = \log \dfrac{z-1}{z+1} = \ln \dfrac{r_1}{r_2} + i(\theta_1 - \theta_2)$

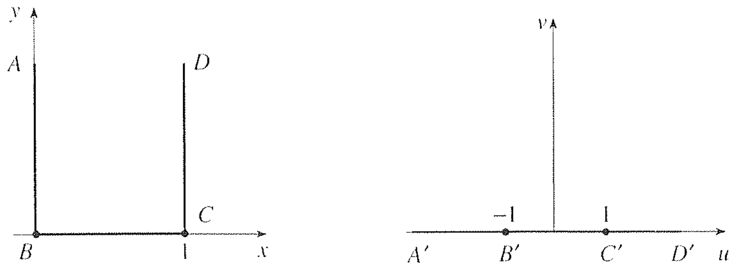


9. $w = \log \dfrac{z-1}{z+1} = \ln \dfrac{r_1}{r_2} + i(\theta_1 - \theta_2)$

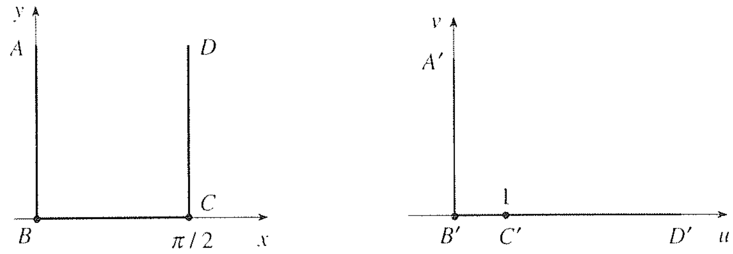


$b = \cos^{-1} \dfrac{a}{\sqrt{a^2+1}}$; $0 < b < \pi$

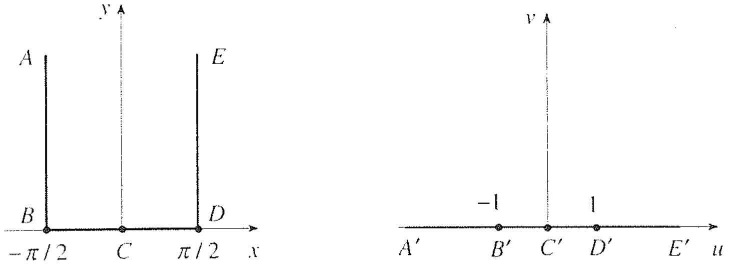
10. $w = -\cos \pi z$



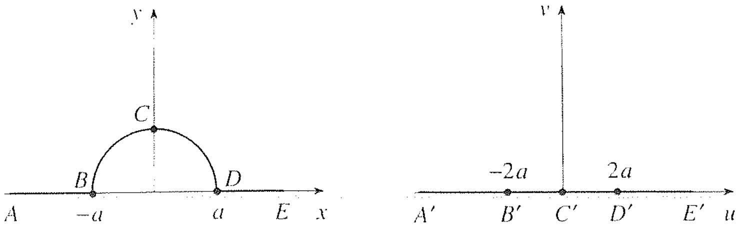
11. $w = \sin z$



12. $w = \sin z$



13. $w = z + \dfrac{a^2}{z}$



Answers to Selected Exercises

Chapter 1

Section 1.2

1. (a) First order, y_1 yes, y_2 no, y_3 no (d) First order, y_1 no, y_2 yes (g) Third order, y_1 yes **5.** (a) -3 (d) 1 (g) ± 1, $\pm\sqrt{5}$ **6.** (a) $A = 0$, $B = 2$ **7.** (a) Linear (d) Nonlinear (because of the $\exp y$ term) (g) Nonlinear (because of the yy''' term)

Section 1.3

2. When $H \to \infty$, with L fixed, the cable hangs straight down at $x = L/2$, and the tension T there is simply half the total weight of the cable, $wL/2$. [In fact, (16) also gives $T(0) = 0$ as $H \to \infty$. Do you see, from the physics, why that result is correct as well?]

Chapter 2

Section 2.2

2. (a) $y = (3x + C)e^x$, $\sigma = e^{-x}$ (d) $\frac{1}{5}(\sin 2x + 2\cos 2x) + Ce^x$, $\sigma = e^{-x}$ (g) $y = x^3 + Cx^2$, $\sigma = x^{-2}$ (for the equation $y' - \frac{2}{x}y = x^2$) (m) $x = t^5 + Ct$, $\sigma = 1/t$ (for the equation $x' - \frac{1}{t}x = 4t^4$). Whereas $p(t) = -1/t$ and $q(t) = 4t^4$ are continuous on $0 < t < \infty$ and $-\infty < t < 0$, the solution $t^5 + Ct$ holds on $-\infty < t < \infty$.
5. (a) $y(x) = 2x^2 - 2/x$ on $0 < x < \infty$ (d) $y(x) = 2x^2$ on $-\infty < x < \infty$
6. (a) $y(x) = (x^3 + 3x^2 - 20)/(3x^2)$ on $-\infty < x < 0$ or on $0 < x < \infty$ (d) $y(x) = (x^3 + 3x^2 - 1)/(3x^2)$ on $-\infty < x < 0$ or on $0 < x < \infty$ **7.** (a) Solution for $y(x)$ given in implicit form by $x = e^y(3y + C)$ **8.** (a) Direction field reveals the straight-line integral curve $y = 2x$ (b) Direction field reveals the straight-line integral curves $y = \pm 2$
9. (a) For $n = 0$, $y = e^{-\int p\,dx}\left(e^{\int p\,dx} q\,dx + C\right)$; for $n = 1$, $y = Ce^{-\int (p-q)\,dx}$ **10.** (a) $y = e^{4x}/(C - e^{4x})$ (d) $y = (x - 2 + Ce^{-x/2})^{2/3}$ (g) $y = -\ln(x - C_1) + C_2$, where both C_1 and C_2 are arbitrary constants. Do you see that this result is equivalent to the solution $y = -\ln(C_1 x + C_2)$, where both C_1 and C_2 are arbitrary constants?
12. (a) $y = 4e^{4x}/(C - e^{4x})$, which does give $y = Y(x) = -4$ for the choice $C = 0$ (d) $y = 2/(Ce^x + e^{-x})$ (g) $y = -2(e^{4x} + C)/(e^{4x} - C)$ **13.** (b) $x(p) = -\frac{3}{2} + \frac{C}{p^2}$, $y(p) = \frac{2C}{p}$ (In this example we can eliminate the parameter p between these equations and obtain $y(x) = \sqrt{2C}\sqrt{2x+3} = A\sqrt{2x+3}$, where $\sqrt{2C} \equiv A$. (f) $P_0 = 1$ gives $y_1(x) = x + e$, $P_0 = 2$ gives $y_2(x) = 2x + e^2$ (i) $P_0 = 0$ gives $y_1(x) = 0$, $P_0 = +2$ gives $y_2(x) = 2x - 2\sin 2$, $P_0 = -2$ gives $y_3(x) = -2x + 2\sin 2$

Section 2.3

1. (a) $t = 4.605L/R$ **2.** (a) For $0 \le t < t_1$, (11) gives $i(t) = i_0 e^{-Rt/L} + \frac{1}{L}\int_0^t e^{R(\tau - t)/L} E_0\,d\tau = \frac{E_0}{R} + \left(i_0 - \frac{E_0}{R}\right)e^{-Rt/L}$. For $t_1 < t < \infty$, (11) gives $i(t) = i_0 e^{-Rt/L} + \frac{1}{L}\int_0^{t_1} e^{R(\tau - t)/L} E_0\,d\tau$

$= \left[i_0 + \frac{E_0}{R}\left(r^{Rt_1/L} - 1\right)\right] e^{-Rt/L}$. Steady state: $i(t) \to 0$ as $t \to \infty$. **3.** (a) If $R^2C \neq L$, $i(t) = [(E_0RC)/(R^2C - L)]\left(e^{-Rt/L} - e^{-t/(RC)}\right)$; if $R^2C = L$, $i(t) = -\frac{E_0}{L}te^{-t/(RC)}$. **5.** 847 years **9.** $N(t) = aN_0/[bN_0 + (a - N_0b)e^{-at}]$ **10.** (a) $c(t) = c_1 + (c_0 - c_1)e^{-Q_0t/v}$ (b) For $0 < t < 1$, $c(t) = e^{-\int_0^t 4\,d\tau}\left(\int_0^t e^{\int_0^\tau 4\,d\xi}4\,d\tau + 0\right) = 1 - e^{-4t}$. Thus, $c(1) = 0.9817$. Then, for $t > 1$, $c(t) = e^{-\int_1^t 2\,d\tau}\left(\int_1^t e^{\int_1^\tau 2\,d\xi}2\,d\tau + 0.9817\right) = 1 - 0.0183e^{-2(t-1)}$. **11.** $x(t) = (mg\sin\alpha/c^2)(ct - m + me^{-ct/m})$ **13.** 0.2299 inches **14.** (a) It is simplest to treat $x < 0$ and $x > 0$ separately. For $x < 0$ we have $Q(x) = 0$ so $c' + (\beta/U)c = 0$ with $c(-\infty) = 0$. Applying $c(-\infty) = 0$ to the general solution $c(x) = Be^{-\beta x/U}$ gives $B = 0$ so $c(x) = 0$ for $x < 0$. For $x > 0$ we have $Q(x) =$ constant $= Q$ and $c' + (\beta/U)c = Q/(AU)$ with $c(0) = 0$ [from the solution $c(x) = 0$ on $x < 0$]. This problem gives $c(x) = \frac{Q}{\beta A}(1 - e^{-\beta x/U})$. Note that $c(x) \to \frac{Q}{\beta A}$ as $x \to \infty$. (b) As in (a), $c(x) = 0$ for $x < 0$ and $c(x) = \frac{Q}{\beta A}(1 - e^{-\beta x/U})$ in $x > 0$, but only up to $x = L$, where it gives $c(L) = \frac{Q}{\beta A}(1 - e^{-\beta L/U})$ as the initial condition for the problem $c' + (\beta/U)c = 0$ on $L < x < \infty$. Applying that initial condition to the general solution $c(x) = Be^{-\beta x/U}$ gives $B = \frac{Q}{\beta A}(e^{\beta L/U} - 1)$ and hence $c(x) = \frac{Q}{\beta A}(e^{\beta(L-x)/U} - e^{-\beta x/U})$ for $x > L$. This time $c(x) \to 0$ as $x \to \infty$.
15. (b) 2.20 hrs

Section 2.4

1. (a) $y = \ln(x^3 + 1)$. We do not claim that these solutions hold for all x. For example, this one holds only on $-1 < x < \infty$ because $\ln(x^3 + 1) \to -\infty$ as $x \to -1$. (d) $y = \tan(x - 0.6266)$ (g) $y = 12e^{3x}/(1 - 4e^{3x})$ (j) $y = -\sqrt{x/3}$ **4.** $N(t) = aN_0/[bN_0 + (a - N_0b)e^{-at}]$ **6.** (a) $y = (1 - \sqrt{8x^3 + 4x + 25})/2$ (d) $y = (1 + \sqrt{8x^3 + 4x + 13})/2$ **7.** (a) $y = -\sqrt{x^3 - x + 9}$ (d) $y = +\sqrt{x^3 - x}$ and $-\sqrt{x^3 - x}$, nonunique solution **8.** (a) No (b) Yes, degree zero. **10.** (a) $y = x(\frac{9}{2}\ln x + C)^{2/3}$ (d) $y = -x + C/x$ **11.** (b) Implicit solution: $y^2 + (6 - 2x)y + 2x^2 - 12x + C = 0$. Can solve for y by quadratic formula.

Section 2.5

1. (a) $y = 3x + C$, $y = 3x + 6$ (d) $10\sin 2u + e^{-5v} = C$, $10\sin 2u + e^{-5v} = e^{30}$ (g) $x^2 + z^2 - 4xz = C$, $x^2 + z^2 - 4xz = 94$ (j) $x^3\sin 2y - x^2y = C$, $x^3\sin 2y - x^2y = -0.7854$ **4.** (a) Exact if and only if $b = A$. **5.** (a) $\sigma(x) = e^{3x}$, $e^{3x}y = C$ (d) $\sigma(y) = e^y$, $xe^y - y = C$ (g) $\sigma(y) = \cos y$, $(x - y)\cos^2 y = C$ (j) $\sigma = 1$, $u^3\sinh 3v - u^2 = C$ **7.** (a) $\sigma = xy$, $x^3y^2 - x^2y^3 = C$ **8.** (a) $\sigma = e^{-(x+y)}$, $e^{-x} + e^{-y} = C$
9. (a) $x^2 - 2xy - y^2 = C$

Chapter 3

Section 3.2

1. (a) No, from Definition 3.2.1. **2.** (a) For instance, $(3x - 5) = 3(x + 2) - 11(1)$. (d) For instance, $\cosh x = -1(\sinh x) + 1(e^x) + 0(e^{2x})$. **3.** (a) The Wronskian determinant $W[1, x, \ldots, x^n](x)$ is upper triangular, with $0!, 1!, 2!, \ldots, n!$ as its diagonal elements. Thus, by property D3 in Section 10.4, $W(x) = (0!)(1!)\cdots(n!)$ which is nonzero. Then, by Theorem 3.2.2 the set is LI (on $-\infty < x < \infty$). An alternative approach that avoids determinants is as follows. Let $n = 2$, say. The question is whether $a + bx + cx^2 = 0$ can be satisfied by constants a, b, c other than $a = b = c = 0$. Repeated differentiation gives $b + 2cx = 0$ and then $2c = 0$, and these three equations give $c = 0$, $b = 0$, $a = 0$ so, by Theorem 3.2.1, $\{1, x, x^2\}$ is LI. The argument is easily generalized to $\{1, x, \ldots, x^n\}$. (d) Surely, e^{2x} is not a scalar multiple of e^x (nor vice versa). For if $e^{2x} = ae^x$ then $x = 0$ gives $a = 1$ and $x = 1$ gives $a = e$, which cannot both be true. Then, by Theorem 3.2.4, the set is LI. **4.** (a) $W = \begin{vmatrix} e^x & e^{2x} & e^{3x} \\ e^x & 2e^{2x} & 3e^{3x} \\ e^x & 4e^{2x} & 9e^{3x} \end{vmatrix} = 2e^{6x} \neq 0$, so LI. Theorem 3.2.4 does not apply. (d) $W = 0$ so, by Theorem 3.2.3, LD.

Section 3.3

1. (a) e^x and e^{2x} satisfy the ODE. Further, $W[e^x, e^{2x}](x) = e^{3x} \neq 0$, so $C_1e^x + C_2e^{2x}$ is a general solution. (d) e^{-x} and e^{2x} satisfy the ODE. Further, $W[e^{-x}, e^{2x}](x) = 3e^x \neq 0$, so $C_1e^{-x} + C_2e^{2x}$ is a general solution. **2.** (a) $e^{3x}, \cosh 3x, \sinh 3x$ are indeed solutions, but they are not LI because $\cosh 3x = 1(\sinh 3x) + 1(e^{-3x})$, for instance. Thus, the set is not a basis. (d) Each is a solution, and $W[e^x, xe^x, x^2e^x](x) = 2e^{3x} \neq 0$ so the set is LI. Thus, the set is a basis. **3.** (a) No; for second-order we need *two* LI solutions. (b) Yes **4.** (a) No; x^2 is not a solution. (c) Yes on $0 < x < \infty$, on $-\infty < x < 0$, and on $6 < x < 10$, but no on $-\infty < x < \infty$ because $x \ln|x|$ is not differentiable at $x = 0$. **5.** (a) No. A general solution of a seventh-order ODE must include seven LI solutions, whereas the given expression contains only six. **7.** No. Note that if y_1 and y_2 are solutions of a homogeneous linear ODE, then $C_1y_1 + C_2y_2$ is a solution too, for arbitrary constants C_1, C_2, but the ODE given here is *not* homogeneous. **8.** (a) Given $y(0) = 4, y'(0) = 3$. Then ODE gives $y'' = -y, y''' = -y', y^{(iv)} = -y'', \dots$ so $y''(0) = -y(0) = -4$, $y'''(0) = -y'(0) = -3$, $y^{(iv)}(0) = -y''(0) = 4$, $y^{(v)}(0) = -y'''(0) = 3$. Hence, $y(x) = 4 + 3x - 4x^2/2! - 3x^3/3! + 4x^4/4! + 3x^5/5! - \cdots$ (d) $y(0) = 1, y'(0) = 0$. ODE gives $y'' = -xy, y''' = -y - xy', y^{(iv)} = -2y' - xy'', y^{(v)} = -3y'' - xy'''$, so $y''(0) = 0, y'''(0) = -1, y^{(iv)}(0) = 0, y^{(v)}(0) = 0$. Hence, $y(x) = 1 + 0x + 0x^2 + (-1)x^3/3! + 0x^4 + 0x^5 + \cdots$ **9.** (a) $p_1(x) = 2$ and $p_2(x) = 3$ are continuous on $-\infty < x < \infty$ so it follows from Theorem 3.3.1 that there exists a unique solution for $y(x)$ on every interval (no matter how broad) containing the initial point $x = 0$. (e) $p_1(x) = 0, p_2(x) = 1, p_3(x) = -1/x$ are continuous on the interval $-\infty < x < 0$ containing the initial point $x = -1$, so there does exist a unique solution on that interval. **11.** (a) Unique solution $y(x) = 0$. (b) Boundary conditions give $C_1 = 0$ and $C_1 = -3$. Hence, no solution. **12.** (a) Unique solution $y(x) = -(2/\pi)x \sin x$.

Section 3.4

3. Trying to avoid the messy Wronskian determinant, fall back on Theorem 3.2.1. Accordingly, the question is whether $\alpha_1e^{\lambda_1x} + \alpha_2xe^{\lambda_1x} + \cdots + \alpha_kx^{k-1}e^{\lambda_1x} = 0$ can be satisfied by α_j's not all zero. Since $e^{\lambda_1x} \neq 0$ for all x, the exponential e^{λ_1x} factors can be cancelled, leaving $\alpha_1 + \alpha_2x + \cdots + \alpha_kx^{k-1} = 0$. Thus, the given set is LI if and only if $\{1, x, \dots, x^{k-1}\}$ is, and proof of the linear independence of that set is given in the answer to Exercise 3(a) of Section 3.2. **4.** (a) $y(x) = A + Be^{-5x}$ (d) $y(x) = Ae^x + Be^{2x}$ is general solution, and initial conditions give $y(x) = 2e^{x-1} - e^{2(x-1)}$ (g) $y(x) = e^{2x}(2\cos x + \sin x)$ (m) $y(x) = Ae^x + e^{-x}(B\cos x + C\sin x)$ **6.** (a) $y(x) = 2 - x$ (d) $y(x) = 1$ (g) $y(x) = Ae^{-x} + (B + Cx)e^x$ **8.** (a) $(\lambda - 2)(\lambda - 6) = \lambda^2 - 8\lambda + 12$, so $y'' - 8y' + 12y = 0$, $y(x) = Ae^{2x} + Be^{6x}$ (d) $y''' - 6y'' - y' + 30y = 0$, $y(x) = Ae^{-2x} + Be^{3x} + Ce^{5x}$ (g) $y^{(v)} - 12y^{(iv)} + 49y''' - 76y'' + 48y' - 64y = 0$, $y(x) = (A + Bx + Cx^2)e^{4x} + D\cos x + E\sin x$ **9.** (a) $y(x) = Ae^{i(1+\sqrt{2})x} + Be^{i(1-\sqrt{2})x} = e^{ix}(C\cos\sqrt{2}\,x + D\sin\sqrt{2}\,x)$ (d) $y(x) = (A + Bx)e^{ix}$ **12.** (a) $\lambda = 0.0776, 1.4612 + 4.8619i, 1.4612 - 4.8619i$, hence unstable (d) $\lambda = +i, -i, -0.5 + 1.9369i, -0.5 - 1.9369i$: in this case there are no roots to the right of the imaginary axis and the two roots *on* the imaginary axis are nonrepeated, hence stable

Section 3.5

1. (a) $E = \sqrt{6^2 + 1^2} = \sqrt{37}$, $\phi = \tan^{-1}(6/1) = 1.406$, $\omega = 1$, so $\sqrt{37}\sin(t + 1.406)$ (d) $\sqrt{8}\sin(3t - 0.785)$ or $\sqrt{8}\sin(3t - \pi/4)$ **3.** $A = x_0$, $B = (cx_0 + 2mx_0')/\sqrt{4m^2\omega^2 - c^2}$ **10.** (d) Yes

Section 3.6

1. (a) $y(x) = C/x$ on $x < 0$ or $x > 0$ (or $-\infty < x < \infty$ if $C = 0$) (d) $y(x) = 3(x^5 - 1)/5$ (g) $y(x) = (2/x)\sin(\ln x)$ on $x > 0$ (p) $y(x) = [A + B\ln|x| + C(\ln|x|)^2]x$ since $\lambda = 1, 1, 1$ **7.** (b) $Y'' - 4Y = 0$, $Y(t) = Ae^{2t} + Be^{-2t}$, $y(x) = Ax^2 + B/x^2$ (f) $Y'' - Y' + Y = 0$, $Y(t) = Ae^{(\frac{1}{2}+i\frac{\sqrt{3}}{2})t} + Be^{(\frac{1}{2}-i\frac{\sqrt{3}}{2})t} = e^{t/2}(C\cos\frac{\sqrt{3}}{2}t + D\sin\frac{\sqrt{3}}{2}t)$, $y(x) = \sqrt{x}\,[C\cos(\frac{\sqrt{3}}{2}\ln|x|) + D\sin(\frac{\sqrt{3}}{2}\ln|x|)]$ on $x < 0$ or $x > 0$ **9.** (a) $\phi(r) = [(\phi_2\ln r_1 - \phi_1\ln r_2) + (\phi_1 - \phi_2)\ln r]/(\ln r_1 - \ln r_2)$ **10.** (a) $u(r) = (u_1r_1 - u_2r_2)/(r_1 - r_2) + [(r_2r_1(u_2 - u_1)/(r_1 - r_2)](1/r)$ **11.** (a) $y(x) = Ax + B\left(e^{-x^2/2} + x\int_0^x e^{-t^2/2}\,dt\right)$ or, in terms of the tabulated function erf (x)

defined by (60), $y(x) = Ax + B[e^{-x^2/2} + \sqrt{\frac{\pi}{2}}x\,\mathrm{erf}\,(\frac{x}{\sqrt{2}})]$. **16.** In that case (50a,b) become $a' = a^2 + (a_1)a + (a_2)$ and $b' = -b^2 - (a_1)b - (a_2)$. Because a_1 and a_2 are constants these two equations admit solutions $a(x) =$ constant and $b(x) =$ constant, which constants are roots of the characteristic equation $\lambda^2 + a_1\lambda + a_2 = 0$. Thus, a, b are the roots λ_1, λ_2 of the characteristic equation, so the factored equation is $(D - \lambda_2)(D - \lambda_1)y = 0$. The latter can be solved by setting $(D - \lambda_1)y \equiv u(x)$. Then $(D - \lambda_2)u = u' - \lambda_2 u = 0$ gives $u(x) = Ae^{\lambda_2 x}$. Finally, solve $(D - \lambda_1)y = u$, namely, $y' - \lambda_1 y = Ae^{\lambda_2 x}$. If $\lambda_1 \neq \lambda_2$, then that step gives $y = C_1e^{\lambda_1 x} + C_2e^{\lambda_2 x}$; if $\lambda_1 = \lambda_2$, then that step gives $y = (C_1 + C_2x)e^{\lambda_1 x}$. **17.** (a) $a_1(x) = -2/x$ and $a_2(x) = 2/x^2$ so (50a) gives the equation $a' = a^2 - \frac{2}{x}a$ on $a(x)$. By inspection, the latter has a particular solution $a(x) = 0$. Rather than try to solve (50b) for $b(x)$, we obtain $b(x)$ readily from (49a) as $b(x) = -a_1(x) - a(x) = 2/x$. Thus, the factored form (47) is $D(D - \frac{2}{x})y = 0$. Then $(D - \frac{2}{x})y = A$ or $y' - \frac{2}{x}y = A$ and (using the general solution of the first-order linear equation) $y(x) = -Ax + Bx^2$ or $C_1x + C_2x^2$. NOTE: Rather than solve (50a) for $a(x)$ and (50b) for $b(x)$, solve (50a) for $a(x)$ and then (49a) for $b(x)$. Or, solve (50b) for $b(x)$ and then (49a) for $a(x)$. In the present example (50a) gives, by inspection, a solution $a(x) = 0$; (50b) gives, by inspection, solutions $b(x) = 1/x$ *and* $b(x) = 2/x$. Of these, the pair $a(x) = 0$ and $b(x) = 2/x$ does satisfy (49a); the pair $a(x) = 0$ and $b(x) = 1/x$ does *not*.

Section 3.7

1. (a) Yes, $\{x^2\cos x, x^2\sin x, x\cos x, x\sin x, \cos x, \sin x, \}$ (b) No, $x^2\ln x \rightarrow \{x^2\ln x, x\ln x, x, 1, 1/x, 1/x^2, \ldots, \}$ without end. **2.** (a) $y_h(x) = C_1e^{3x}$, $xe^{2x} \rightarrow \{xe^{2x}, e^{2x}\}$, $6 \rightarrow \{1\}$. Seek $y_p(x) = Axe^{2x} + Be^{2x}$ for the xe^{2x} term and $y_p(x) = C$ for the 6 term. Obtain $A = -1$, $B = -1$, $C = -2$, so $y(x) = C_1e^{3x} - xe^{2x} - e^{2x} - 2$.
(d) $y_h(x) = C_1e^{3x}$, $xe^{3x} \rightarrow \{xe^{3x}, e^{3x}\}$, $4 \rightarrow \{1\}$. Seek $y_p(x) = x(Axe^{3x} + Be^{3x})$ for the xe^{3x} term and $y_p(x) = C$ for the 4 term. Obtain $y(x) = C_1e^{3x} + \frac{1}{2}x^2e^{3x} - \frac{4}{3}$. (g) $y(x) = C_1 + C_2e^x + \frac{1}{2}\cos 2x - \sin 2x$
(k) $y(x) = C_1\cos x + C_2\sin x + 3x\sin x + 2$ (o) $y(x) = C_1 + C_2e^x + (x^2 - 2x)e^x$ **4.** (a) $y(x) = C_1e^{-2x} + e^{2x}$
(d) $y(x) = C_1\frac{1}{x} + \frac{\ln x}{x}$ (g) $y(x) = C_1e^x + C_2e^{-x} + (4x - 2)e^x$ (m) $y(x) = C_1x^2 + C_2/x^2 - 1/(3x)$
(o) $y(x) = C_1e^x + (C_2 + C_3x)e^{-x} + 1 - x$

Section 3.8

5. (a) As $c \rightarrow \infty$ the graph of E becomes discontinuous at $\Omega = 0$: $E = F_0/k$ at $\Omega = 0$, $E = 0$ for all $\Omega > 0$.
(b) For $c = 0$, Φ is 0 for $\Omega < \omega$ and π for $\Omega > \omega$; for $c \rightarrow \infty$, Φ is 0 at $\Omega = 0$ and $\pi/2$ for $\Omega > 0$.
8. Of course we can see from (19b) that $E \rightarrow 0$ as $\Omega \rightarrow \infty$. One way of interpreting this result physically is to let $\Omega t = \tau$ in (15), in which case (15) becomes $m\Omega^2d^2x/d\tau^2 + c\Omega dx/d\tau + kx = F_0\cos\tau$ in which the "effective mass" $m\Omega^2$ tends to ∞ as $\Omega \rightarrow \infty$ (as does the effective damping $c\Omega$). **12.** (b) (12.2) is $mw'' + cw' + kw = F_0e^{i\Omega t}$ and seeking $w_p(t) = Ae^{i\Omega t}$ gives $A = F_0/(-m\Omega^2 + ic\Omega + k)$ so $x_p(t) = \mathrm{Re}\,[F_0e^{i\Omega t}/(-m\Omega^2 + ic\Omega + k)] = F_0[(k - m\Omega^2)\cos\Omega t + c\Omega\sin\Omega t]/[(k - m\Omega^2)^2 + c^2\Omega^2]$ (e) $x_p(t) = \mathrm{Im}\,[e^{i3t}/(3i - 1)] = -\frac{6}{5}\cos 3t - \frac{2}{5}\sin 3t$
13. (a) $Q(t) = 5 - (5/3)e^{-t}(3\cos 3t + \sin 3t)$. Steady state: $Q(t) \rightarrow 5$.

Section 3.9

4. (c) $R(i_1 - i_2) = E(t)$, $L(i_2' - i_3') = E(t)$, $(1/C)i_3 = E'(t)$, of which the first and third happen to be algebraic rather than differential equations. **5.** (a) $x(t) = -2Ae^{2t} - Be^{-t}$, $y(t) = Ae^{2t} + 2Be^{-t}$ (d) $x(t) = Ae^{t/2} - 3Be^{-t/2} - 1$, $y(t) = Ae^{t/2} + Be^{-t/2} - t - 1$ (g) $x(t) = -\frac{13}{16} + \frac{1}{8}t + 3Ae^{4t} + Be^{-4t}$, $y(t) = -\frac{7}{8} + \frac{1}{8}t + Ae^{4t} - Be^{-4t}$
(k) $x(t) = -2Ae^{at} - 2Be^{-at} + C\sin at + D\cos at$, $y(t) = Ae^{at} + Be^{-at} + C\sin at + D\cos at$ $(a = \sqrt{3})$ **8.** Let us give only $z(t)$:
(a) $z(t) = \gamma + \gamma(\beta e^{-\alpha t} - \alpha e^{-\beta t})/(\alpha - \beta)$ (b) $z(t) = \gamma(1 - e^{-\alpha t} - \alpha te^{-\alpha t})$ **10.** (a) Given $(D - 1)x + y = t$, $(D^2 - 1)x + (D + 1)y = t^2$, operate on the first with $D + 1$ and obtain $(D^2 - 1)x + (D + 1)y = 1 + t$. Since $t^2 \neq 1 + t$ there is no solution. (b) Operating on the first equation with $D + 1$ gives an equation that is identical to the second equation. Thus, we can discard the second equation. In the first equation we can set $x(t) = f(t)$, an arbitrary (twice-differentiable) function and solve, by algebra, for $y(t)$. Doing so, we obtain $y(t) = t - f'(t) + f(t)$.

Chapter 4

Section 4.2

1. (a) $a_n = n$, $\lim |a_{n+1}/a_n| = \lim |(n+1)/n| = 1$ so $R = 1$ (d) $a_n = 0$ for all $n > 1000$, so the "series" converges for all x; i.e., it consists of only a finite number of terms so convergence is not an issue. $R = \infty$
(g) $\lim |a_{n+1}/a_n| = \lim |(\frac{n+1}{n})^{50}\frac{1}{n}| = 0$ so $R = \infty$ **2.** (a) Denominator has zeros at $\pm i$ in z plane so $R = 1$.
(d) Denominator vanishes at $x = -2$ so $R = 24$. (g) Factoring gives $(x+2)(x-1)/[(x^2+4)(x-1)] = (x+2)/[(x+2i)(x-2i)]$. Distance, in z plane, from 2 to $\pm 2i$ is $\sqrt{8}$, so $R = \sqrt{8}$. **3.** (a) $e^x = e\sum_0^\infty (x-1)^n/n!$, $R = \infty$ (d) $\sin x = 1 - \frac{1}{2!}(x - \frac{\pi}{2})^2 + \frac{1}{4!}(x - \frac{\pi}{2})^4 - \cdots = \sum_0^\infty \frac{(-1)^n}{(2n)!}(x - \frac{\pi}{2})^{2n}$, $R = \infty$ (g) $\cos x = \cos 5 - \frac{\sin 5}{1!}(x-5) - \frac{\cos 5}{2!}(x-5)^2 + \frac{\sin 5}{3!}(x-5)^3 + \frac{\cos 5}{4!}(x-5)^4 - \cdots$, $\lim |a_{n+1}/a_n| = 0$, $R = \infty$ (j) $2x^3 - 4 = -4 + 0x + 0x^2 + \frac{12}{3!}x^3 + 0x^4 + \cdots = -4 + 2x^3$, $R = 0$ **5.** (c) At $x = 1$ the geometric series is $1 + 1 + 1 + \cdots$, which diverges because the nth term does not tend to zero as $n \to \infty$ (see the italics below Theorem 4.2.1). Alternatively, $s_n = n \to \infty$ as $n \to \infty$, so the series diverges. At $x = -1$ the series is $1 - 1 + 1 - 1 + \cdots$, which diverges because the nth term does not tend to zero. **7.** (a) $p(x) = 2$, $q(x) = 1$, $R = \infty$. $y(x) = \sum_0^\infty a_n x^n$ gives the recursion formula $(n+2)(n+1)a_{n+2} + 2(n+1)a_{n+1} + a_n = 0$ for $n = 0, 1, 2, \ldots$. Obtain $y(x) = a_0 + a_1 x - (\frac{a_0}{2} + a_1)x^2 + (\frac{a_0}{3} + \frac{a_1}{2})x^3 - (\frac{a_0}{8} + \frac{a_1}{6})x^4 + \cdots$ so $y_1(x) = 1 - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{8}x^4 + \cdots$, $y_2(x) = x - x^2 + \frac{1}{2}x^3 - \frac{1}{6}x^4 + \cdots$. NOTE: Since the analytical solution is $y(x) = (A + Bx)e^{-x}$ we might expect y_1 and y_2 to be e^{-x} and xe^{-x}. Actually, y_2 *is* xe^{-x}, but y_1 is not e^{-x}; it is $(1+x)e^{-x}$. That's fine; $y_1 = (1+x)e^{-x}$ and $y_2 = xe^{-x}$ are, indeed, two LI solutions of the ODE, LI because neither is a scalar multiple of the other.
(d) $p(x) = 1/x$, $q(x) = 1/x$, $R = 5$. Don't forget to expand the x in the ODE as $-5 + (x+5)$. $y(x) = \sum_0^\infty a_n(x+5)^n$ gives the recursion formula $-5(n+2)(n+1)a_{n+2} + (n+1)^2 a_{n+1} + a_n = 0$ for $n = 0, 1, 2, \ldots$. Obtain $y(x) = a_0 + a_1(x+5) + (\frac{a_0}{10} + \frac{a_1}{10})(x+5)^2 + (\frac{a_0}{75} + \frac{7a_1}{150})(x+5)^3 + (\frac{11a_0}{3000} + \frac{13a_1}{1500})(x+5)^4 + \cdots$ so $y_1(x) = 1 + \frac{1}{10}(x+5)^2 + \frac{1}{75}(x+5)^3 + \frac{11}{3000}(x+5)^4 + \cdots$, $y_2(x) = (x+5) + \frac{1}{10}(x+5)^2 + \frac{7}{150}(x+5)^3 + \frac{13}{1500}(x+5)^4 + \cdots$.
(g) $p(x) = (3+x)/x$, $q(x) = 1$, $R = 3$. Recursion formula is $-3(n+2)(n+1)a_{n+2} + n(n+1)a_{n+1} + (n-3)a_n + a_{n-1} = 0$ for $n = 0, 1, 2, \ldots$, with $a_{-1} = 0$. Obtain $y_1(x) = 1 - \frac{1}{2}(x+3)^2 + \frac{1}{72}(x+3)^4 - \frac{1}{180}(x+3)^5 + \cdots$, $y_2(x) = (x+3) - \frac{1}{9}(x+3)^3 + \frac{1}{108}(x+3)^4 + \frac{1}{540}(x+3)^5 + \cdots$. **11.** (a) $a_{n+2} = \frac{(n+1)^2}{(n+3)(n+2)}a_{n+1} - \frac{n}{(n+3)(n+2)}a_n \sim a_{n+1} - \frac{1}{n}a_n$ so $\lim |a_{n+2}/a_{n+1}| = 1$, $R = 1$ (d) Because of the subscripts $n+2$ and n, successive terms in $y_1(x)$ and $y_2(x)$ differ by a factor of $(x - x_0)^2$ rather than $(x - x_0)$. Thus, the ratio test requires $\lim_{n\to\infty} |a_{n+2}(x - x_0)^{n+2}/[a_n(x - x_0)^n]| = \lim_{n\to\infty} \left|\frac{a_{n+2}}{a_n}\right| (x - x_0)^2 < 1$ for convergence. If $\lim |a_{n+2}/a_n| \equiv L$, then we have convergence in $|x - x_0| < 1/\sqrt{L}$. In the present case $a_{n+2}/a_n = 3(n+2)/(n+1) \to 3 = L$, so $R = 1/\sqrt{3}$.

Section 4.3

1. (a) $p = -x^3$ and $q(x) = x$ are infinitely differentiable and hence analytic (recall our italicized rule of thumb in the sentence preceding Section 4.2.2) for all x so there are no singular points; every point is an ordinary point.
(d) $p(x) = 0$ and $q(x) = 1/[x(x^2+3)]$ are analytic for all x except $x = 0$, which is a singular point, a regular singular point because $xp(x) = 0$ and $x^2 q(x) = x/(x^2+3)$ are analytic there. (g) $p(x) = q(x) = (x-1)^{-1}(x+3)^{-2}$. Singular points at $x = 1$, $x = -3$. $(x-1)p(x) = (x+3)^{-2}$ and $(x-1)^2 q(x) = (x-1)(x+3)^{-2}$ are analytic at $x = 1$, so $x = 1$ is a regular singular point. $(x+3)p(x) = (x-1)/(x+3)$ and $(x+3)^2 q(x) = 1/(x-1)$. The latter is analytic at $x = -3$ but the former is not, so $x = -3$ is an irregular singular point. **3.** (a) $tY'' + Y' - Y = 0$
5. (a) The indicial equation is (38): $r^2 + (p_0 - 1)r + q_0 = 0$. If $r = 1$ and 4 then $(r-1)(r-4) = r^2 - 5r + 4 = 0$ so $p_0 - 1 = 5$ and $q_0 = 4$; hence $p_0 = -4$, $q_0 = 4$. Hence, $xp(x) = p_0 + p_1 x + \cdots = -4$ and $x^2 q(x) = q_0 + q_1 x + \cdots = 4$, say, so $p(x) = -4/x$, $q(x) = 4/x^2$, and the ODE is $y'' - (4/x)y' + (4/x^2)y = 0$ or $x^2 y'' - 4xy' + 4y = 0$. Or we could take $xp(x) = -4 + 3x - 5x^3$ and $x^2 q(x) = 4 + x^2$, or $xp(x) = -4\cos x$ and $x^2 q(x) = 4e^x$, and so on. (d) $r^2 + (p_0 - 1)r + q_0 = 0$ and $(r + 1/2)(r - 1/2) = r^2 - 1/4 = 0$ give $p_0 = 1$, $q_0 = -1/4$. Can choose $xp(x) = 1$ and $x^2 q(x) = -1/4$, for instance. Thus, $x^2 y'' + xy' - \frac{1}{4}y = 0$, say. **6.** (a) $p(x) = 1/(2x)$ is singular

at $x = 0$, but $xp(x) = 1/2$ and $x^2q(x) = x^4/2$ are analytic there (with radii of convergence $R_1 = R_2 = \infty$), so $x = 0$ is a regular singular point; $p_0 = 1/2$ and $q_0 = 0$ so indicial equation is $r^2 - \frac{1}{2}r = 0$; $r = 0, 1/2$; case (i). $y_1(x) = 1 - \frac{1}{28}x^4 + \frac{1}{3360}x^8 - \cdots$, $y_2(x) = \sqrt{x}\left(1 - \frac{1}{36}x^4 + \frac{1}{4896}x^8 - \cdots\right)$, both valid on $0 < x < \infty$.
(d) $p(x) = 1/x$ is singular at $x = 0$, but $xp(x) = 1$ and $x^2q(x) = x^2$ are analytic there (with $R_1 = R_2 = \infty$), so $x = 0$ is a regular singular point; $p_0 = 1$ and $q_0 = 0$ so indicial equation is $r^2 = 0$; $r = 0, 0$; case (ii). $y_1(x) = 1 - \frac{1}{2^2}x^2 + \frac{1}{2^24^2}x^4 - \frac{1}{2^24^26^2}x^6 + \cdots = \sum_0^\infty (-1)^n x^{2n}/[\,1^2 \cdot 2^2 \cdot 4^2 \cdots (2n)^2\,]$ and (41b), with $r = 0$, gives $y_2(x) = y_1(x)\ln x + (\frac{1}{4}x^2 - \frac{3}{128}x^4 + \frac{11}{13824}x^6 - \cdots)$, both valid on $0 < x < \infty$. (g) $p(x) = 1/x$ and $q(x) = -(1 + 2x)/x^2$ are singular at $x = 0$, but $xp(x) = 1$ and $x^2q(x) = -1 - 2x$ are analytic there (with $R_1 = R_2 = \infty$), so $x = 0$ is a regular singular point; $p_0 = 1$ and $q_0 = -1$ so indicial equation is $r^2 - 1 = 0$; $r = -1, 1$; case (iii). $y_1(x) = x + \frac{2}{3}x^2 + \frac{1}{6}x^3 + \cdots$, $y_2(x) = 4y_1(x)\ln x + \frac{1}{x}(-2 + 4x - \frac{32}{9}x^3 + \cdots)$, both valid on $0 < x < \infty$.

Section 4.4

10. (a) It is convenient to first shift the origin to the point of expansion, $x = 1$, by setting $x - 1 \equiv t$. Then the ODE becomes $t(2 + t)Y'' + 2(1 + t)Y' - 2Y = 0$ on $Y(t)$. The indicial equation is $r^2 = 0$ so $r = 0, 0$; case (ii) of Theorem 4.3.1. Obtain one solution $Y_1(t) = 1 + t = x$ [i.e., the bounded solution $P_1(x)$] and the second solution $Y_2(t) = (\ln t)(1 + t) + \left(-\frac{5}{2}t - \frac{3}{8}t^2 + \frac{1}{12}t^3 + \cdots\right)$. Observe that the $\ln t$ term reveals the singular nature of that solution as $t \to 0$ [i.e., as $x \to 1$ since $\ln t = \ln(1 - x)$]. Similarly, at $x = -1$ we obtain one solution x and a second solution with a $\ln(1 + x)$ term, which "blows up" as $x \to -1$.

Section 4.5

3. (a) $\int_0^\infty dx/(x^4 + 2) = \int_0^1 dx/(x^4 + 2) + \int_1^\infty dx/(x^4 + 2)$. $\int_0^1$ converges because $1/(x^4 + 2)$ is continuous on $0 \le x \le 1$, and $\int_1^\infty$ converges by Theorem 4.5.2 (b) because $1/(x^4 + 2) \sim 1/x^4$ as $x \to \infty$ and $\int_1^\infty dx/x^4$ is a convergent p-integral (Theorem 4.5.1). Observe that it was *necessary* to break up the integral, for example as $\int_0^1 + \int_1^\infty$, because otherwise our I_2 integral (in Theorem 4.5.2) would be $\int_0^\infty dx/x^4$ and $1/x^4$ is not bounded on $0 \le x < \infty$, as assumed in the theorem. (d) Break up $\int_0^\infty = \int_0^1 + \int_1^\infty$. $\int_0^1$ converges because the integrand is continuous on $0 \le x \le 1$, but $\int_1^\infty$ diverges [Theorem 4.5.2 (b)] because $x^{3.2}/(x^4 + 100) \sim 1/x^{0.8}$ and $\int_1^\infty dx/x^{0.8}$ is a divergent horizontal p-integral (Theorem 4.5.1). Thus, the given integral diverges. (g) The integral is positive on $4 \le x < \infty$ and $\frac{\sin^2 x}{\sqrt{x}(x-1)} \le \frac{1}{\sqrt{x}(x-1)} \sim \frac{1}{x^{3/2}}$. Since $\int_4^\infty dx/x^{3/2}$ converges (Theorem 4.5.1 with $p = 3/2 > 1$), the given integral converges as well [Theorem 4.5.2 (b) and (a)]. (j) $1/(x^2 \cos x) \sim 1/x^2$ as $x \to 0$ so, by Theorem 4.5.5, the integral diverges. **6.** (a) Obtain $I = \int_0^{1/2} \xi^2\, d\xi/(1+2\xi^4)$, which integral is not singular; the limits are finite and the integrand is continuous on $0 \le \xi \le 1/2$, so I converges. Thus, we may be able to "desingularize" a convergent singular integral by a suitable change of variables. This idea is especially important regarding the *numerical* evaluation of integrals. For instance, it would be much easier to evaluate the regular ξ integral (above) than the original singular integral, because of the *finite* extent of the ξ integration interval. **7.** (a) $1 < \alpha < \infty$ (c) As $x \to \infty$, $x^\alpha/(x + 1) \sim 1/x^{1-\alpha}$ so we need $1 - \alpha > 1$ or $\alpha < 0$; as $x \to 0$, $x^\alpha/(x + 1) \sim x^\alpha = 1/x^{-\alpha}$ so we need $-\alpha < 1$ or $\alpha > -1$. Thus, for convergence we need $-1 < \alpha < 0$. (e) $I = \int_1^2 (x - 1)^\alpha (x + 1)^\alpha\, dx$ so the integrand blows up at $x = 1$ if $\alpha < 0$. By vertical p-integral test, we need $\alpha > -1$ for convergence. If this is not clear, let $x - 1 = t$, to move the singularity to the origin. Then $I = \int_0^1 t^\alpha (t + 2)^\alpha\, dt$ and $t^\alpha(t + 2)^\alpha \sim 2^\alpha/t^{-\alpha}$ as $t \to 0$, so we need $-\alpha < 1$ or $\alpha > -1$.
8. (a) (16) gives $\Gamma(3.5) = 2.5\Gamma(2.5) = (2.5)(1.5)\Gamma(1.5) = (2.5)(1.5)(0.5)\Gamma(0.5) = (2.5)(1.5)(0.5)\sqrt{\pi} = 3.323$, which appears to be consistent with Fig. 3. (b) (22) gives $\Gamma(-3.5) = \Gamma(-2.5)/(-3.5) = \Gamma(-1.5)/(-3.5)(-2.5) = \Gamma(-0.5)/(-3.5)(-2.5)(-1.5) = \Gamma(0.5)/(-3.5)(-2.5)(-1.5)(-0.5) = \sqrt{\pi}/(3.5)(2.5)(1.5)(0.5) = 0.270$, which appears to be consistent with Fig. 3. **10.** (a) With $x^p = t$, $I = \frac{1}{p}\int_0^\infty e^{-t}t^{(1/p)-1}\, dt = \frac{1}{p}\Gamma(\frac{1}{p})$. (c) Let $x^2 = t$. **13.** HINT: Sketch the graph of $\exp(-x^p)$, on $0 \le x < \infty$, as $p \to \infty$. **19.** (a) $F(x) \sim 4 = O(1)$ as $x \to 0$

(d) $G(x) \sim -3/4 = O(1)$ as $x \to \infty$ (h) $I(x) \sim 2x/x^2 = 2/x = O(1/x)$ as $x \to \infty$

Section 4.6

12. (a) Comparing $y'' + 4x^2y = 0$ with $y'' + (a/x)y' + bx^{c-a}y = 0$ gives $a = 0$, $b = 4$, $c = 2$ so $\alpha = 1/2$, $\nu = 1/4$ and $y(x) = x^{1/2}Z_{1/4}(\frac{1}{2}\sqrt{4}x^2)$, $y(x) = A\sqrt{x}J_{1/4}(x^2) + B\sqrt{x}Y_{1/4}(x^2)$. (d) Comparing $y'' + (9/4)xy = 0$ with $y'' + (a/x)y' + bx^{c-a}y = 0$ gives $a = 0$, $b = 9/4$, $c = 1$ so $\alpha = 2/3$, $\nu = 1/3$ and $y(x) = x^{1/2}Z_{1/3}(\frac{2}{3}\sqrt{\frac{9}{4}}x^{3/2})$, $y(x) = A\sqrt{x}J_{1/3}(x^{3/2}) + B\sqrt{x}Y_{1/3}(x^{3/2})$. (g) $a = 3$, $b = -1$, $c = 3$ so $\alpha = 1$, $\nu = -1$ and $y(x) = x^{-1}Z_{-1}(\sqrt{|-1|}x) = x^{-1}Z_1(x)$, $y(x) = Ax^{-1}I_1(x) + Bx^{-1}K_1(x)$. NOTE: $b = -1 < 0$ so Z_1 denotes I and K, not J and Y. [See sentence following (50).] (j) $a = 3$, $b = -4$, $c = 3$ so $\alpha = 1$, $\nu = -1$ and $y(x) = x^{-1}Z_{-1}(\sqrt{|-4|}x) = x^{-1}Z_1(2x)$, $y(x) = Ax^{-1}I_1(2x) + Bx^{-1}K_1(2x)$. **13.** (a) $a = 3$, $b = 9$, $c = 3$ so $\alpha = 1$, $\nu = -1$ and $y(x) = Ax^{-1}J_1(3x) + Bx^{-1}Y_1(3x)$. (b) Because $x^{-1} \to \infty$, $J_1(3x) \to 0$, and $Y_1(3x) \to -\infty$ as $x \to 0$, we need to be careful in applying the initial conditions. Using (16b) we have $y(x) = \frac{A}{x}\left(\frac{3x}{2} - \frac{27x^3}{16} + \cdots\right) + \frac{B}{x}Y_1(3x) = A\left(\frac{3}{2} - \frac{27x^2}{16} + \cdots\right) + \frac{B}{x}Y_1(3x)$. Since both $1/x \to \infty$ and $Y_1(3x) \to -\infty$ as $x \to 0$, we need to set $B = 0$ if we are to satisfy the condition $y(0) = 6$. Then $y(0) = 6 = 3A/2$ gives $A = 4$. Further, $y'(x) = A(-\frac{27}{8}x + \cdots)$ does happen to satisfy the other initial condition $y'(0) = 0$, so $y(x) = 4x^{-1}J_1(3x)$. (c) We saw that the condition $y(0) = 6$ implied both that $B = 0$ and that $A = 4$, giving $y(x) = 4x^{-1}J_1(3x)$. The latter gives $y'(0) = 0$ so there is no solution satisfying the initial conditions $y(0) = 6$, $y'(0) = 2$. This result does not contradict Theorem 3.3.1 because that theorem supposes that $p_1(x), \ldots, p_n(x)$ are continuous on the closed interval, whereas in this example $p_1(x) = 3/x$ is not continuous at the left endpoint $x = 0$.

Chapter 5

Section 5.2

1. (a) Yes; $K = 5$ (or greater), $c = 4$ (or greater), $T = 0$ (or greater). (d) Yes; $|\cosh 3t| = (e^{3t} + e^{-3t})/2 \le e^{3t}$ for all $t \ge 0$, so we can take $K = 1$, $c = 3$, $T = 0$. (g) Yes; $|\cos t^3| \le 1 = 1e^{0t}$ so we can take $K = 1$, $c = 0$, $T = 0$. (j) No; $e^{t^4}/e^{ct} = e^{t^4-ct} \sim e^{t^4} \to \infty$ as $t \to \infty$, so there do not exist constants K, c, T such that $\exp(t^4) \le Ke^{ct}$ for all $t \ge T$.

Section 5.3

1. (a) $3/[s(s+8)] = \frac{3}{8}\frac{1}{s} - \frac{3}{8}\frac{1}{s+8}$ so linearity and entries 1 and 2 give $f(t) = (3/8)(1 - e^{-8t})$. Alternatively, $1/s \to 1$ and $1/(s+8) \to e^{-8t}$ so the convolution theorem (and linearity) give $3/[s(s+8)] = 3 * e^{-8t} = 3\int_0^t e^{-8\tau}\,d\tau = (3/8)(1 - e^{-8t})$. (d) $5(e^{-2t/3} - e^{-t})$ **3.** (a) Choose $a = -4$, $b = 4i$ (or $-4i$) in entry 9 and obtain $e^{-4t}\sin(4it)/4i = e^{-4t}(e^{i(4it)} - e^{-i(4it)})/[(2i)(4i)] = (1 - e^{-8t})/8$. (d) Choose $a = 1/2$, $b = 3i/2$ in entry 9 and obtain $(e^{2t} - e^{-t})/3$. **10.** (a) $f(t) = \int_0^t e^{t-\tau}\sin 2\tau\,d\tau = e^t * \sin 2t$ so $L\{f(t)\} = L\{e^t\}L\{\sin 2t\} = \frac{1}{s-1}\frac{2}{s^2+4}$. (d) $f(t) = 1 * \cosh t$ so $L\{f(t)\} = \frac{1}{s}\frac{s}{s^2-1} = \frac{1}{s^2-1}$. **11.** (a) $F(s)G(s) = \frac{1}{s^2}\frac{1}{s-1}$ and $L^{-1}\{\frac{1}{s^2(s-1)}\} = e^t - t - 1 \ne te^t$.

Section 5.4

1. (a) $x(t) = 2t^2 - 2t + 1 + x(0)e^{-2t}$ (d) $x(t) = 2 - t + t^3$ (g) $x(t) = 5e^t - 2e^{2t}$ (j) $x(t) = [1 - (1 + 21t)e^{-3t}]/9$ (m) $x(t) = \frac{1}{6}t^3 + \frac{1}{4}t^2 - \frac{1}{4}t + \frac{5}{8} + \frac{1}{56}e^{t/2}[21\cos(\frac{\sqrt{7}t}{2}) + \sqrt{7}\sin(\frac{\sqrt{7}t}{2})]$ (s) $x(t) = 2t^2 - \frac{1}{3}t + \frac{1}{9}(1 - e^{-3t})$

Section 5.5

1. (a) $f(t) = t[H(t) - H(t-2)] + (4-t)[H(t-2) - H(t-4)] = tH(t) + (4-2t)H(t-2) - (4-t)H(t-4)$, $F(s) = \frac{1}{s} + (e^{-4s} - 2e^{-2s})/s^2$ (d) $f(t) = (t^3 - t)[H(t) - H(t-1)] - 6H(t-1)$, $F(s) = \frac{6}{s^4} - \frac{1}{s^2} - 2\frac{e^{-s}}{s^4}(3 + 3s + s^2 + 3s^3)$

3. (a) $-t-10$ NOTE: Recall that $H(t)=1$ for $t>0$. (d) $(t-a)H(t-a)-(t-b)H(t-b)$ (g) $(5-t)H(5-t)$ (j) $\int_0^t e^{-(t-\tau)}H(\tau-5)\,d\tau=0$ for $t<5$ and $e^{-t}\int_5^t e^{\tau}\,d\tau=1-e^{5-t}$ for $t>5$. Thus, we obtain $(1-e^{5-t})H(t-5)$. **5.** (a) $X(s)=e^{-s}/[s(s-1)]$, $x(t)=(e^{t-1}-1)H(t-1)$ (d) $f(t)=10[H(t-5)-H(t-7)]$, $x(t)=10(e^{t-5}-1)H(t-5)-10(e^{t-7}-1)H(t-7)$ (h) $f(t)=t[1-H(t-1)]+(2-t)[H(t-1)-H(t-2)]$, $X(s)=1/[s^2(s-1)]-2e^{-s}/[s^2(s-1)]+e^{-2s}/[s^2(s-1)]$, $x(t)=-t-1+e^t+2(t-e^{t-1})H(t-1)+(1-t+e^{t-2})H(t-2)$

Section 5.6

1. (a) $X(s)=e^{-2s}/(s^2-1)$, $x(t)=H(t-2)\sinh(t-2)$ by entries 30 and 5. (d) $x(t)=t+2(1-e^{-t})+H(t-2)(1-e^{-(t-2)})$ (g) $x(t)=8e^t-4e^{2t}+100H(t-3)(e^{2(t-3)}-e^{t-3})$ **2.** (b) Letting $\kappa t=\tau$, $\int_{-\infty}^{\infty}g(t)\delta(\kappa t)\,dt=\int_{-\infty}^{\infty}g(\tau/\kappa)\delta(\tau)\,d\tau/\kappa=g(0)/\kappa$ if $\kappa>0$ and $\int_{\infty}^{-\infty}g(\tau/\kappa)\delta(\tau)\,d\tau/\kappa=-g(0)/\kappa$ if $\kappa<0$. Thus, $\int_{-\infty}^{\infty}g(t)\delta(\kappa t)\,dt=-g(0)/|\kappa|$. Similarly, $\int_{-\infty}^{\infty}g(t)\frac{\delta(t)}{|\kappa|}\,dt=g(0)/|\kappa|$, so $\delta(\kappa t)=\delta(t)/|\kappa|$. (c) $\int_{-\infty}^{\infty}g(t)[f(t)\delta(t)]\,dt=\int_{-\infty}^{\infty}[g(t)f(t)]\delta(t)]\,dt=g(0)f(0)$. Also, $\int_{-\infty}^{\infty}g(t)[f(0)\delta(t)]\,dt=f(0)\int_{-\infty}^{\infty}g(t)\delta(t)\,dt=f(0)g(0)$, so $f(t)\delta(t)=f(0)\delta(t)$, for $f(0)\neq 0$. If $f(0)=0$, then $\int_{-\infty}^{\infty}g(t)[f(t)\delta(t)]\,dt=\int_{-\infty}^{\infty}[g(t)f(t)]\delta(t)]\,dt=g(0)f(0)=0$. Also, $\int_{-\infty}^{\infty}g(t)[0]\,dt=0$, so $0\delta(t)=0$. (d) $\int_{-\infty}^{t}\delta(\tau)\,d\tau=0$ if $t<0$ and 1 if $t>0$, which is $H(t)$.

Section 5.7

1. (a) $L^{-1}\{\frac{1}{s}\frac{s}{(s^2+a^2)^2}\}=f(t)$ so $f(t)=1*\frac{t\sin at}{2a}$ (entry 11) so $f(t)=\frac{1}{2a}\int_0^t\tau\sin a\tau\,d\tau=(\sin at-at\cos at)/(2a^3)$. $sF(s)\to 0$ as $s\to\infty$, and $f(0)$ does equal 0. (d) Taylor expanding s^2 about $s=-1$ gives $1-2(s+1)+(s+1)^2$, so $s^2/(s+1)^3=1/(s+1)^3-2/(s+1)^2+1/(s+1)$ and, using entries 1, 7, 29, $f(t)=e^{-t}\frac{t^2}{2}-2e^{-t}t+e^{-t}$. $sF(s)\to 1$ as $s\to\infty$ and $f(0)$ does equal 1. (g) $1/s^5\to t^4/4!=t^4/24$ by entry 7 and $e^{-s}/s^5\to H(t-1)(t-1)^4/24$ by entry 30. $sF(s)\to 0$ as $s\to\infty$ and $f(0)$ does equal 0. (j) $F(s)=\ln(s^2+a^2)-\ln s^2$, $dF/ds=\frac{2s}{s^2+a^2}-\frac{2}{s}\to 2\cos at-2$ (entries 4 and 1) so, by entry 31, $-tf(t)=2\cos at-2$, $f(t)=2(1-\cos at)/t$. $sF(s)\to s\ln(1+a^2/s^2)\to(\infty)(0)$ as $s\to 0$. For l'Hôpital we need $0/0$ or ∞/∞, so write $s\ln(1+a^2/s^2)=\frac{\ln(1+a^2/s^2)}{1/s}\to 0$ as $s\to\infty$ by l'Hôpital and, also by l'Hôpital, $f(t)\to 0$ as $t\to 0$. (m) $1/(s^2+s+1)=1/[(s+\frac{1}{2})^2+\frac{3}{4}]\to\frac{e^{-t/2}\sin\sqrt{3/4}\,t}{\sqrt{3/4}}$ (entries 3, 29), so $e^{-s}/(s^2+s+1)\to\frac{2}{\sqrt{3}}H(t-1)e^{-(t-1)/2}\sin 2\sqrt{3}(t-1)$ (entry 30). $sF(s)\to 0$ as $s\to\infty$ and $f(0)$ does equal 0. (p) Proceed as in Example 5. $\frac{d}{ds}\ln\left(\frac{s^2+1}{s^2+s}\right)=\frac{2s}{s^2+1}-\frac{2s+1}{s^2+s}$ so $\ln\left(\frac{s^2+1}{s^2+s}\right)=-\int_s^{\infty}\left(\frac{2\sigma}{\sigma^2+1}-\frac{2\sigma+1}{\sigma^2+\sigma}\right)d\sigma$, "$f(t)$" $=L^{-1}\left\{-\frac{2s}{s^2+1}+\frac{2s+1}{s^2+s}\right\}=-2\cos t+1+e^{-t}$, so entry 32 gives $f(t)=$ "$f(t)$"$/t=(-2\cos t+1+e^{-t})/t$. $sF(s)\to 0$ as $s\to\infty$ and l'Hôpital does give $f(t)\to 0$ as $t\to 0$. (s) $\frac{1}{s}\tanh s=\frac{1}{s}\frac{e^s-e^{-s}}{e^s+e^{-s}}=\frac{1}{s}\frac{1-e^{-2s}}{1+e^{-2s}}=\frac{1}{s}(1-e^{-2s})(1-e^{-2s}+e^{-4s}-e^{-6s}+\cdots)=\frac{1}{s}(1-2e^{-2s}+2e^{-4s}-2e^{-6s}+\cdots)\to H(t)-2H(t-2)+2H(t-4)-H(t-6)+\cdots$, the graph of which is seen to be a square wave $f(t)=1$ on $0<t<2$, -1 on $2<t<4$, and of period 4. As a check, we can verify that the right-hand side of (37) does give us back $\frac{1}{s}\tanh s$. $sF(s)\to 1$ as $s\to\infty$ and we do have $f(t)\to 1$ as $t\to 0$. **5.** (a) $F(s)=\frac{1}{1-e^{-\pi s}}\int_0^{\pi}(\sin t)e^{-st}\,dt=\frac{1+e^{-\pi s}}{1-e^{-\pi s}}\frac{1}{s^2+1}=\coth(\pi s/2)/(s^2+1)$ **8.** $x(t)=x_0\cos t+x_0'\sin t+f(t)*\sin t=x_0\cos t+x_0'\sin t+\int_0^t[2-4H(\tau-2)+4H(\tau-4)-4H(\tau-6)-\cdots]\sin(t-\tau)\,d\tau=x_0\cos t+x_0'\sin t+2t-4H(t-2)[1-\cos(t-2)]+4H(t-4)[1-\cos(t-4)]-4H(t-6)[1-\cos(t-6)]+\cdots$. With $x_0=x_0'=1$, $x(5)=3.2036$.

Chapter 6

Section 6.2

2. (a) $y_1=0.8, y_2=0.64, y_3=0.512$ (d) $y_1=-0.2, y_2=0.0192, y_3=0.2194$ (g) $y_1=y_2=y_3=0$. Indeed, the solution is $y(x)=0$. **3.** (g) Exact solution is $y(x)=4e^x-x^2-2x-3$. Sample results: $y_5=2.253091, y_{10}=$

4.534344; $y(x_5) = y(0.5) = 2.344885$, $y(x_{10}) = y(1) = 4.873127$, so $E_{10} = 0.338783$.

4. (a) For $h = 0.1, 0.05, 0.01, 0.005, 0.001$ Euler gives the values $1.2, 1.225, 1.245, 1.2475, 1.2495$, respectively. Since the exact value is $y(0.5) = 1.25$, the accumulated truncation error is $0.05, 0.025, 0.005, 0.0025, 0.0005$, respectively. These results are consistent with the method being a first-order method, i.e., with the accumulated truncation error being proportional to the step size h, because each time we reduce h the error reduces proportionately. (d) For $h = 0.1, 0.05, 0.01, 0.005, 0.001$ Euler gives the values $1.101786, 1.115792, 1.127303, 1.128761, 1.129932$. The solution is $y(x) = \exp(1 - \cos x)$, so the exact value is $y(0.5) = 1.130226$, hence the accumulated truncation error is $0.02844, 0.014434, 0.002923, 0.001465, 0.000294$, which does indeed diminish proportional to h (approximately).

Section 6.3

1. (a) Second-order: $y_1 = 2.15, y_2 = 2.486476$. Fourth-order: $y_1 = 2.140149, y_2 = 2.477369$. Exact: $y(x) = (4500x^2 + 8)^{1/3}$, so $y(x_1) = y(0.02) = 2.13997$, $y(x_2) = y(0.04) = 2.47712$. (d) Second-order: $y_1 = -7.749214, y_2 = -7.495329$. Fourth-order: $y_1 = -7.749231, y_2 = -7.495363$. Exact: $y(x) = -8\cos x/\cos 1$, so $y(x_1) = y(1.02) = -7.749231$, $y(x_2) = y(1.04) = -7.495363$. **2.** (a) Second-order: $y_{10} = 10.68520$. Fourth-order: $y_{10} = 10.44818$. Exact: $y(x) = (4500x^2 + 1)^{1/3}$, so $y(x_{10}) = y(0.5) = 10.40350$. (d) Second-order: $y_{10} = 0.877527$. Fourth-order: $y_{10} = 0.877583$. Exact: $y(x) = \cos x$, so $y(x_{10}) = y(0.5) = 0.877583$.

3. (a) $x = 1$: Euler gives $y_{10} = 3.434368$ for $h = 0.1$ and $y_{20} = 3.570633$ for $h = 0.05$. Exact $y(1) = 3.718282$ so (28) gives $p \approx 0.943$. $x = 2$: Euler gives $y_{20} = 10.154750$ for $h = 0.1$ and $y_{40} = 10.737989$ for $h = 0.05$. Exact $y(2) = 11.389056$ so (28) gives $p \approx 0.923$. (b) $x = 1$: Second-order R-K gives $y_{10} = 3.705918$ for $h = 0.1$ and $y_{20} = 3.715097$ for $h = 0.05$. Exact $y(1) = 3.718282$ so (28) gives $p \approx 1.957$. $x = 2$: Second-order R-K gives $y_{20} = 11.335919$ for $h = 0.1$ and $y_{40} = 11.3753429$ for $h = 0.05$. Exact $y(2) = 11.389056$ so (28) gives $p \approx 1.954$

4. (a) $x(0) = 0$, $x(600) = 3.314$, $x(1200) = 3.852$, $x(1800) = 3.967$, $x(2400) = 3.993$, $x(3000) = 3.998$, $x(3600) = 4.000$ (d) $x(0) = 6$, $x(600) = 4.487$, $x(1200) = 4.111$, $x(1800) = 4.025$, $x(2400) = 4.006$, $x(3000) = 4.001$, $x(3600) = 4.000$ NOTE: Setting $x'(t) = 0$ in the differential equation gives $0 = 0.02 - 0.01\sqrt{x}$ so the steady-state solution is $x \to 4$. **7.** (b) $C = 0.09824$, $x(600) = 3.313860967, \ldots, x(3600) = 3.999636784$

10. (a) $h = 0.05$ gives 3.718281474, $h = 0.02$ gives 3.718281819, exact $y(1) = 3.718281828$ so (28) gives $p \approx 4.008$. (b) $h = 0.05$ gives 3.701259429, $h = 0.02$ gives 3.711558039 so (28) gives $p \approx 1.01$. Evidently, the error has reduced the method to a first-order method. **13.** (a) Exact $y(x) = 1/(1 - x^2)$ so $y_1 = y(0.1) = \underline{1.01010}$, $y_2 = y(0.2) = \underline{1.04167}$, $y_3 = y(0.3) = \underline{1.09890}$. Then, A-B predictor (31a) gives $y_4 = y_3 + [55(2)(0.3)y_3^2 - 59(2)(0.2)y_2^2 + 37(2)(0.1)y_1^2 - 9(2)(0)y_0^2](0.1/24) = 1.18970$ and taking this as $y_4^{(0)}$, A-M corrector (33) gives $y_4^{(1)} = y_3 + [9(2)(0.4)y_4^{(0)\,2} + 19(2)(0.3)y_3^2 - 5(2)(0.2)y_2^2 + (2)(0.1)y_1^2](0.1/24) = 1.09890 + (7.2y_4^{(0)\,2} + 11.80033)(0.1/24) = 1.19053$, $y_4^{(2)} = 1.09890 + (7.2y_4^{(1)\,2} + 11.80033)(0.1/24) = 1.19059$, $y_4^{(3)} = 1.09890 + (7.2y_4^{(2)\,2} + 11.80033)(0.1/24) = \underline{1.19059}$. (b) Using the four underlined values in (a), the A-B predictor gives $y_5 = y_4 + [55(2)(0.4)y_4^2 - 59(2)(0.3)y_3^2 + 37(2)(0.2)y_2^2 - 9(2)(0.1)y_1^2](0.1/24) = 1.33161$, which we use as $y_5^{(0)}$ in the A-M corrector formula $y_5^{(1)} = y_4 + [9(2)(0.5)y_5^{(0)\,2} + 19(2)(0.4)y_4^2 - 5(2)(0.3)y_3^2 + (2)(0.2)y_2^2](0.1/24) = 1.19059 + (9y_5^{(0)\,2} + 18.35736)(0.1/24) = 1.33357$, $y_5^{(2)} = 1.19059 + (9y_5^{(1)\,2} + 18.35736)(0.1/24) = 1.33377$, $y_5^{(3)} = 1.19059 + (9y_5^{(2)\,2} + 18.35736)(0.1/24) = \underline{1.33379}$.

Section 6.4

2. (a) Exact: $y(x_1) = 0.9800666$, $z(x_1) = -0.1986693$. Euler: $y_1 = 1$, $z_1 = -0.2$. Second-order RK: $y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$, $z_{n+1} = z_n + \frac{1}{2}(l_1 + l_2)$. With $n = 0$, $k_1 = hf(x_n, y_n, z_n) = hz_n = hz_0 = 0$, $l_1 = hg(x_n, y_n, z_n) = -hy_n = -hy_0 = -0.2$, $k_2 = hf(x_{n+1}, y_n + k_1, z_n + l_1) = h(z_0 + l_1) = -0.04$, $l_2 = hg(x_{n+1}, y_n + k_1, z_n + l_1) = -h(y_0 + k_1) = -0.2$, so $y_1 = 1 + \frac{1}{2}(0 - 0.04) = 0.98$, $z_1 = 0 + \frac{1}{2}(-0.2 - 0.2) = -0.2$. Fourth-order RK: $y_0 = 1$, $z_0 = 0$, $k_1 = 0$, $l_1 = -0.2$, $k_2 = -0.02$, $l_2 = -0.2$, $k_3 = -0.02$, $l_3 = -0.198$, $k_4 = -0.0396$, $l_4 = -0.196$, so $y_1 = 0.9800667$, $z_1 = -0.1986667$ (d) Exact: $y(x_1) = 1.44$, $z(x_1) = 1.2$. Euler: $y_1 = 1.4$, $z_1 = 1.2$. Second-order RK: $k_1 = 0.4$, $l_1 = 0.2$, $k_2 = 0.49371$, $l_2 = 0.19444$, $y_1 = 1.4468$, $z_1 = 1.1972$. Fourth-order RK: $k_1 = 0.4$, $l_1 = 0.2$, $k_2 = 0.4436667$, $l_2 = 0.1983471$, $k_3 = 0.4350841$, $l_3 = 0.2022597$, $k_4 = 0.4834599$,

$l_4 = 0.1985687$, so $y_1 = 1.4401603$, $z_1 = 1.1999637$ **5.** (a) $x' = u$, $x(0) = x_0$; $u' = [f(t) - kx - cu]/m$, $u(0) = x_0'$ (d) $y' = u$, $y(0) = 2$; $u' = 3x + 4y - u$, $u(0) = 1$ **6.** (a) $y(1) = 1.0418532996$, $z(1) = -1.3328381583$, $y(2) = 4.4563094187$, $z(2) = -0.9372059061$ **8.** (a) $L[Y_1] = 0$, $Y_1(0) = 1$, $Y_1'(0) = 0$; $L[Y_2] = 0$, $Y_2(0) = 0$, $Y_2'(0) = 1$; $L[Y_p] = 3\sin x$, $Y_p(0) = Y_p'(0) = 0$. Then $y(x) = C_1Y_1(x) + C_2Y_2(x) + Y_p(x)$. $y(0) = 1$ gives $C_1 = 1$ and $y(2) = 3$ gives $C_2 = [3 - Y_1(2) - Y_p(2)]/Y_2(2)$. Using *Maple* dsolve command with abserr = Float (1, −6) (to obtain the desired accuracy) gives $Y_1(1) = 0.338991$, $Y_2(1) = 1.219163$, $Y_p(1) = 0.626700$, $Y_1(2) = -8.366992$, $Y_2(2) = 7.639154$, $Y_p(2) = 14.523026$. Thus, $C_2 = -0.413139$. Hence, $y(1) = 0.46201$.

Section 6.5

1. (a) No. $y(x) = Ce^{2x} + 4x$ and $y(0) = 0$ gives $C = 0$ so $y(x) = 4x$. In effect, numerical error causes C to differ slightly from 0, in which case the e^{2x} term will cause the solution to diverge dramatically from the exact solution $4x$. **6.** (a) $\partial f/\partial y = e^{x+y} > 0$ on $0 \le x \le 4$; stable (b) $\partial f/\partial y = -e^{x-y} < 0$ on $0 \le x \le 4$; unstable **11.** (a) General solution $y_n = A(4^n)$; particular solution $y_n = 5(4^n)$. (d) General solution $y_n = A + B(3^n)$; particular solution $y_n = 4 - 3^n$. **15.** (a) Particular solution $Y_n = -\frac{1}{2}n - \frac{1}{4}$; general solution $y_n = -\frac{1}{2}n - \frac{1}{4} + A(3^n)$. (d) Particular solution $Y_n = n^2$; general solution $y_n = n^2 + A(2^n) + B(3^n)$. (g) Particular solution $Y_n = e^n/(e^2 - 1)$; general solution $y_n = e^n/(e^2 - 1) + A + B(-1)^n$.

Chapter 7

Section 7.2

4. (a) $x^2 + y^2 = C^2$; concentric circles; flow clockwise. **5.** (a) Trajectories are the straight lines $y = -x + C$; rightward flow for $y > 0$, leftward for $y < 0$; singular points all along the x axis. (d) $x^2 - (y/3)^2 = C^2$; equilibrium point at origin.
6. Use the *Maple* command

```
phaseportrait ([y, -x + x^3], [t, x, y], t = 0..20, {[0, .05, 0], [0, .3, 0], [0, .6, 0], [0, .9, 0], [0, .95, 0], [0, .99, 0]},
stepsize = .05, scene = [t, x]);
```

The periods are (approximately, from the plot) 6.3, 6.7, 7.5, 10.5, 12, 17, respectively. They do approach 2π as $A \to 0$ and show signs of tending to infinity as $A \to 1$.

9. (a) $y^2 + x^4 = C$. **11.** Equilibrium points at $(0, 0)$ and $(1, 1)$. The axes are trajectories, with the flow downward on the y axis and rightward on the x axis. Within the first quadrant the trajectories are closed orbits containing the equilibrium point $(1, 1)$, although the lineal element field does not really permit us to distinguish between periodic orbits and weak spirals, by eye.

Section 7.3

1. (a) $0 = 0$ and $2x - y = 0$ give the *line* of singular points $y = 2x$; hence, not isolated. (d) Isolated singular points at $(-n\pi, n\pi)$ for $n = 0, \pm1, \pm2, \ldots$. (g) Isolated singular points at $(2\sqrt{2}, \sqrt{2})$ and $(-2\sqrt{2}, -\sqrt{2})$. **4.** No; for a given ϵ the corresponding δ could be enormous **9.** (a) Saddle; $y = 2x$ (unstable manifold), $y = -2x$ (stable manifold) (d) Saddle; $y = x/\sqrt{3}$ (unstable), $y = -x/\sqrt{3}$ (stable) (g) Unstable node; $y = x$ (unstable), $y = -x$ (unstable) **11.** (a) Unstable focus (d) Center (g) Unstable node

Section 7.4

1.(b) If we write $x'' + \epsilon x'^3 + x = 0$ as $x'' + (\epsilon x'^2)x' + x = 0$ we can think of $(\epsilon x'^2)$ as the (variable) damping coefficient. For large motions x'^2 is large over much of the motion so the effective damping is greater than for the linear case $x'' + \epsilon x' + x = 0$. But as the motion diminishes x'^2 becomes small so the effective damping tends to zero. This result should be observable from the phaseportrait in part (a). **2.**(a) Saddle point at $(-1, 0)$, center at (1,0) (d) Singular points at $x = y = n\pi/2$ $(n = 0, \pm1, \pm2, \ldots)$, saddles for n odd and unstable foci for n

even. (g) Borderline case: stable focus, stable proper node, or stable improper node at (0,0); a phaseportrait plot reveals it to be a stable improper node. (j) Stable focus at (0,0), saddle at (4,8). **4.** (a) No, because of the x' term. (c) Yes **6.** (a) If $r = 0.3$, (32) gives $S_+ = (x_+, y_+) = (3, 0.9)$ and $S_- = (x_-, y_-) = (1/3, 0.1)$. At S_+ obtain $X' = -0.3X + Y, Y' = 0.06X - Y$, a stable improper node with straight-line trajectories $Y = -0.7772X$ and $Y = 0.0772X$. At S_- obtain $X' = -0.3X + Y, Y' = 0.54X - Y$, a saddle with straight-line trajectories $Y = -1.16394X$ (stable manifold) and $Y = 0.46394X$ (unstable manifold). At the origin obtain $x' = -0.3x + y$, $y' = 0x - y$, a stable node with straight- line trajectories $y = 0$ and $y = -0.7x$. **8.** (a) Singular points at $x_\pm = (1 \pm \sqrt{(1-4r)})/2, y = 0$, where $0 < x_- < x_+ < 1$. $(x_+, 0)$ is a saddle and $(x_-, 0)$ is a center or a focus, but we can rule out a focus because the system is conservative. Trajectories given by $y^2 + x^2 + 2r \ln|x - 1| =$ constant= C. Of these, the separatrix is $y^2 + x^2 + 2r \ln|x-1| = x_+^2 + 2r\ln(1 - x_+)$. At the right, the vertical line $x = 1$ is a trajectory, with upward flow direction. [As $r \to 1/4$, $x_\pm$ merge at $x = 1/2$ and the merging center and saddle produce a higher-order singularity with no periodic motions, and for $r > 1/4$ the singularity disappears altogether. Physically, for $r > 1/4$ the force of attraction is so great as to rule out the possibility of oscillation and the vertical line $x = 1$ is an symptote for every trajectory. These points are covered in parts (c)–(g).]

Section 7.5

2. (a) $f(x) = x^2 - 1, g(x) = x^3$. f is even; g is odd; $g(x) > 0$ for all $x > 0$; $g'(x) = 3x^2$ is continuous for all x. $F(x) = \int_0^x f(\zeta)d\zeta = x^3/3 - x$ so, with $x_0 = \sqrt{3}$, $F(x) < 0$ for all $0 < x < x_0$; $F(x) > 0$ for $x > x_0$; $F(x)$ is monotone increasing for $x > x_0$; $F(x) \to \infty$ as $x \to \infty$. Thus, there exists a single limit cycle enclosing the origin.

Section 7.6

4. Saddle at $(0, 0)$; stable foci at $(1, 0)$ and $(-1, 0)$ if $r < \sqrt{8}$, stable nodes if $r > \sqrt{8}$.

Chapter 8

Section 8.2

1. (a) No; e.g., $6x^2 - x + 3 = 0$ (d) Yes (g) A linear equation cannot be transcendental

Section 8.3

1. Final results:
(a) $x = 7/17, y = -1/17$ (unique solution)
(d) $z = \alpha, y = 6 + 3\alpha, x = 7 + 2\alpha$ (nonunique solution: a 1– parameter family of solutions)
(g) $z = \alpha, y = 3 - 2\alpha, x = -2 + \alpha$ (nonunique solution: a 1– parameter family of solutions)
(j) $x_4 = \alpha, x_3 = \beta, x_2 = 2 + \alpha + \beta, x_1 = 1 - \beta$ (nonunique solution: a 2–parameter family of solutions)
4. Yes; yes; yes; yes; no. At most we can have a 14- parameter family of solutions, and that will occur if and only if *all* of the a_{ij}'s and c_j's in (1) are zero. **7.** (a) The equations are homogeneous because we can write them as $(2 - \lambda)x + y = 0, x + (2 - \lambda)y = 0$. There is only the unique trivial solution $x = y = 0$ unless $\lambda = 1$ or $\lambda = 3$. If $\lambda = 1$ there is the 1– parameter family of solutions $y = \alpha$, $x = -\alpha$ (which contains, but is not limited to the trivial solution). If $\lambda = 3$ there is the 1–parameter family of solutions $y = \alpha, x = \alpha$.
(d) By inspection, if $\lambda = 0$ there is the 2–parameter family of solutions $x = \alpha$, $y = \beta$, $z = 0$, and if $\lambda \neq 0$, there is the 1– parameter family of solutions $x = y = \alpha$, $z = \lambda\alpha$. **8.** (a) Slow down: if we add -2 times the first equation to the second, as a replacement for the second we obtain $x_1 - 2x_2 = 0$ and $0 = 0$. Now if we add $-1/2$ times the second to the first we still have $x_1 - 2x_2 = 0$ and $0 = 0$. **10.** Don't physical systems such as this always have unique solutions? For any choice of values of R_1, R_2, R_3 the system (10.1) will indeed have a *unique* solution for i_1, i_2, i_3, except for these cases: Case 1. If $R_2 = R_3 = 0$ there is a *nonunique* solution: $i_1 = E/R_1$ and i_2 and i_3 sum to i_1, but are not uniquely determined ($i_3 = \alpha, i_2 = E/R_1 - \alpha$). Case 2. If R_1 and either R_2 or R_3 are zero, then there is *no solution*: the system is *inconsistent*. If $R_1 = R_3 = 0$, for example, then the last equation is $0 = E$, which has no

solutions. More physically, if $R_1 = R_3 = 0$, then we have a "short circuit" and the currents will be infinite. That is what the last equation ($0i_1 + 0i_2 + 0i_3 = E$) is trying to tell us for, crudely speaking, zero times something can equal a nonzero number only if that something is infinite.

Chapter 9

Section 9.3

1. (a) -25. **4.** (a) $\mathbf{OC} \cdot \mathbf{AB} = (\mathbf{OA} + \mathbf{AD} + \mathbf{DC}) \cdot \mathbf{AB} = \mathbf{OA} \cdot \mathbf{AB} + \mathbf{AD} \cdot \mathbf{AB} + \mathbf{DC} \cdot \mathbf{AB} = (1)(1)\cos 90^o + (1)(1)\cos 90^o + (1)(1)\cos 0^o = 1$. By expressing $\mathbf{OC}$ as $\mathbf{OA} + \mathbf{AD} + \mathbf{DC}$ and using (3.1) we obtain several simple dot products, simple because θ is 0^o or 90^o. (d) $\mathbf{OC} \cdot \mathbf{CP} = (\mathbf{OA} + \mathbf{AD} + \mathbf{DC}) \cdot (\frac{1}{2}\mathbf{CD} + \frac{1}{2}\mathbf{AO}) = -1$. NOTE: There is no name for the midpoint of CD so for the vector from that point to P we have used $\frac{1}{2}\mathbf{AO}$.
5. (a) $APO = \cos^{-1}[(\mathbf{PA} \cdot \mathbf{PO})/(\|\mathbf{PA}\|\|\mathbf{PO}\|)]$, where $\mathbf{PA} \cdot \mathbf{PO} = (\frac{1}{2}\mathbf{OA} + \frac{1}{2}\mathbf{CD} + \mathbf{DA}) \cdot (\frac{1}{2}\mathbf{AO} + \frac{1}{2}\mathbf{CD} + \mathbf{DA}) =$ etc. $= 1$; $\|\mathbf{PA}\| = \sqrt{\mathbf{PA} \cdot \mathbf{PA}} = \sqrt{(\frac{1}{2}\mathbf{OA} + \frac{1}{2}\mathbf{CD} + \mathbf{DA}) \cdot (\frac{1}{2}\mathbf{OA} + \frac{1}{2}\mathbf{CD} + \mathbf{DA})} = \sqrt{6}/2$, $\|\mathbf{PO}\| =$ etc. $= \sqrt{6}/2$, so $APO = \cos^{-1}(2/3) = 48.19^o$.

Section 9.4

1. (a) $(24, -7, 23, 32)$ (d) Not defined because of the $\mathbf{tu}$ product. NOTE: Actually, one can extend our mathematical system to include products of vectors, which are called **dyads** or **second order tensors**, but even if we do that we cannot add the dyad $4\mathbf{tu}$ to the vector $\mathbf{w}$. **2.** (a) $\mathbf{x} = [-5/9, 1, 20/9, -2/3]$ **4.** Only the trivial solution exists: $\alpha_1 = \alpha_2 = \alpha_3 = 0$.

Section 9.5

1. (a) $\|\mathbf{u}\| = 5$, $\|\mathbf{v}\| = \sqrt{5}$, $\theta = 63.4^o$ (d) $\|\mathbf{u}\| = 2\sqrt{3}$, $\|\mathbf{v}\| = \sqrt{77}$ **2.** (a) Yes, it is a scalar times a vector. (b) No, it is a vector dotted with a scalar. **3.** (a) $ABC = \cos^{-1}(-1/\sqrt{10}) = 108.4^o$, $BCA = \cos^{-1}(2/\sqrt{5}) = 26.6^o$, $CAB = \cos^{-1}(1/\sqrt{2}) = 45^o$. **4.** (a) $\hat{\mathbf{u}} = [4/5, 3/5]$, $\hat{\mathbf{v}} = [2/\sqrt{5}, -1/\sqrt{5}]$ **6.** (a) $\mathbf{u}_1 = \mathbf{u} = [1, 3, 0]$, $\mathbf{u}_2 = \mathbf{u} - \frac{10}{11}\mathbf{v} = [-9/11, 3/11, 0]$, $\mathbf{u}_3 = \mathbf{u} - \frac{5}{4}\mathbf{v} + \frac{3}{4}\mathbf{w} = [0, 0, -9/4]$ **7.** (a) $3\sqrt{3}$ **9.** (a) $\mathbf{u} \cdot [3, 0, -1] = 3u_1 - u_3 = 0$, so $u_3 = \alpha$, $u_2 = \beta$, $u_1 = \alpha/3$ and $\mathbf{u} = [\alpha/3, \beta, \alpha]$ (α, β arbitrary) (d) $\mathbf{u} = [-15\alpha - 4\beta, 5\alpha - 8\beta, 7\beta, 7\alpha]$ (α, β arbitrary) **10.** (c) $\mathbf{u}_1 = [0, 18/13, 27/13]$, $\mathbf{u}_2 = [2, 21/13, -14/13]$ **12.** (b) $l_1 = 2/\sqrt{30}$, $l_2 = -1/\sqrt{30}$, $l_3 = 5/\sqrt{30}$ **14.** (a) Yes **15.** (a) Gauss elimination gives $x_3 = 3\alpha$, say, $x_2 = 8/3 + 2\alpha$, $x_1 = 8/3 - \alpha$, so $\mathbf{x} = [8/3 - \alpha, 8/3 + 2\alpha, 3\alpha] = [8/3, 8/3, 0] + \alpha[-1, 2, 3]$ so the desired vector is $\pm\frac{1}{\sqrt{14}}[-1, 2, 3]$.

Section 9.6

1. (a) Yes (f) No; (6) is not satisfied

Section 9.7

1. (a) Yes (d) No (g) No **3.** (a) No: there is no negative inverse $-\mathbf{u}$ for each $\mathbf{u}$ in the set. Also, the set is not closed under scalar multiplication because $\alpha\mathbf{u}$ is not in the set if $\alpha < 0$. **4.** (a) span$\{[-4, 0, 1], [1, 1, 0]\}$ (d) span$\{[-1, 0, 0, 1], [-2, 1, 1, 0]\}$ **5.** (a) $[-4, 0, 1]$, $[2, 1, 0]$ **6.** (a) Yes (b) No **7.** $[1, 2]$, $[2, -1]$

Section 9.8

1. (a) No: a set is one or the other. **2.** (a) $[3, 4] = 2[1, 1] + [1, 2]$ **3.** (a) LD: e.g., $[7, 3] = [1, 3] + 3[2, 0] + 0[-1, 3]$ (d) LD: e.g., $[2, 3, 0] = -\frac{1}{3}[1, -2, 4] + [1, 1, 0] + \frac{4}{3}[1, 1, 1]$

Section 9.9

1. (a) No (d) No (g) Yes (o) Yes **2.** (a) $[9,-2,4]=2\mathbf{e}_1+\frac{13}{5}\mathbf{e}_2+\frac{2}{5}\mathbf{e}_3$ **3.** (a) $[9,-2,4]=2\sqrt{14}\,\hat{\mathbf{e}}_1+\frac{13}{\sqrt{5}}\hat{\mathbf{e}}_2+\frac{2\sqrt{70}}{5}\hat{\mathbf{e}}_3$ **4.** (a) $[1,0,0,0]=\frac{1}{15}\mathbf{e}_1+\frac{1}{3}\mathbf{e}_2+0\mathbf{e}_3+\frac{1}{5}\mathbf{e}_4$ **8.** (a) 1 (d) 3 **9.** (a) 3 (d) 1 **10.** (a) 2 **12.** (a) [1,0], [0,1] (d) $\frac{1}{\sqrt{2}}[1,1,0],\ \frac{1}{\sqrt{22}}[3,-3,2],\ \frac{1}{\sqrt{11}}[-1,1,3]$ **13.** (b) $\mathbf{e}_1^*=[1,-1],\ \mathbf{e}_2^*=[0,1],\ \mathbf{u}=(\mathbf{u}\cdot\mathbf{e}_1^*)\mathbf{e}_1+(\mathbf{u}\cdot\mathbf{e}_2^*)\mathbf{e}_2=2\mathbf{e}_1+\mathbf{e}_2$ (e) $\mathbf{e}_1^*=[1,-1,0],\ \mathbf{e}_2^*=[0,1,-1],\ \mathbf{e}_3^*=[0,0,1],\ \mathbf{u}=5\mathbf{e}_1-6\mathbf{e}_2+5\mathbf{e}_3,\ \mathbf{v}=0\mathbf{e}_1-2\mathbf{e}_2+2\mathbf{e}_3,\ \mathbf{w}=7\mathbf{e}_1-5\mathbf{e}_2+3\mathbf{e}_3$

Section 9.10

2. $\mathbf{u}\approx[64/15,0,-19/30,0,11/6],\ \|\mathbf{E}\|=4.0042$ **3.** In $\text{span}\{\hat{\mathbf{e}}_1\},\mathbf{u}\approx-\frac{4}{3}[1,1,0,-1]$ with $\|\mathbf{E}\|=7.188$; in $\text{span}\{\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2\},\mathbf{u}\approx\frac{1}{3}[1,-9,-5,4]$ with $\|\mathbf{E}\|=6.236$; in $\text{span}\{\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3\},\mathbf{u}\approx[4,-3,2,5]$ with $\|\mathbf{E}\|=1.732$; in $\text{span}\{\hat{\mathbf{e}}_1,\hat{\mathbf{e}}_2,\hat{\mathbf{e}}_3,\hat{\mathbf{e}}_4\},\mathbf{u}=[4,-2,1,6]$ with $\|\mathbf{E}\|=0$

Chapter 10

Section 10.2

2. (a) undefined (d) 6×1 (g) 4×1 **4.** No; $\mathbf{A}-c$ is undefined because $\mathbf{A}$ is $n\times n$ and c is a scalar.
5. (a) $(\mathbf{A}+\mathbf{B})(\mathbf{A}+\mathbf{B})=\mathbf{A}^2+\mathbf{AB}+\mathbf{BA}+\mathbf{B}^2=\mathbf{A}^2+2\mathbf{AB}+\mathbf{B}^2$ only if $\mathbf{AB}=\mathbf{BA}$, that is, if $\mathbf{A}$ and $\mathbf{B}$ commute; in *general* $\mathbf{AB}\neq\mathbf{BA}$ so the given formula is incorrect. **7.** (a) $2\mathbf{A}^2+4\mathbf{AB}+\mathbf{BA}+2\mathbf{B}^2$
8. (a) $\mathbf{A}^{100}$ is a 2×2 matrix with each element equal to 2^{99}. (d) $\mathbf{D}^{100}$ is a 3×3 zero matrix.
10. (a) Dimensionally, $[m\times n][4\times1]=1\times1$ so $n=4$ and $m=1$. Thus, $\mathbf{A}$ is a 1×4 matrix, i.e., a single row $[a_{11},\ a_{12},\ a_{13},\ a_{14}]$. Multiplying the latter into $\mathbf{x}=[x_1,x_2,x_3,x_4,]^{\mathrm{T}}$ gives $a_{11}x_1+a_{12}x_2+a_{13}x_3+a_{14}x_4$ and comparing the latter with x_1-3x_4 shows that $a_{11}=1,\ a_{12}=0,\ a_{13}=0,\ a_{14}=-3$, so $\mathbf{A}=[1,0,0,-3]$.
11. (a) For example, $\mathbf{AB}=\begin{bmatrix}3&0\\4&0\end{bmatrix}\begin{bmatrix}0&0\\6&5\end{bmatrix}=\begin{bmatrix}0&0\\0&0\end{bmatrix}=\mathbf{0}$ **16.** (c) $a_{11}=\alpha$ (arbitrary), $a_{21}=\beta$ (arbitrary), $a_{12}=(1-\alpha^2)/\beta,\ a_{22}=-\alpha$ or $a_{11}=\pm1,\ a_{12}=a_{21}=0,\ a_{22}=\pm1$ with the signs of a_{11} and a_{22} being independent of each other.

Section 10.3

1. (a) $\mathbf{x}^{\mathrm{T}}\mathbf{y}=[-3],\ \mathbf{xy}^{\mathrm{T}}=\begin{bmatrix}3&6\\-3&-6\end{bmatrix}$ **7.** $\mathbf{A}=\begin{bmatrix}3&1.5\\1.5&-5\end{bmatrix}+\begin{bmatrix}0&0.5\\-0.5&0\end{bmatrix}=$ symmetric + skew symmetric **8.** (a) $\mathbf{A}=\begin{bmatrix}6&-4\\-4&1\end{bmatrix}$

Section 10.4

2. (a) 0 (g) 132 **14.** (a) a_j's >0, $\Delta_1=6$, $\Delta_2=\begin{vmatrix}6&1\\4&5\end{vmatrix}=26>0$, $\Delta_3=\begin{vmatrix}6&1&0\\4&5&6\\0&1&4\end{vmatrix}=68>0$,
$\Delta_4=\begin{vmatrix}6&1&0&0\\4&5&6&1\\0&1&4&5\\0&0&0&1\end{vmatrix}=68>0$; stable. As a check, the *Maple* solution, using fsolve, gives $\lambda=-5.18,-0.34,-0.24\ \pm0.72i$, each of which has a negative real part.

Section 10.5

1. (a) $r=1$, nullity $=3$, number of LI rows $=1$, number of LI columns $=1$ (e) $r=2$, nullity $=1$, number of LI rows $=2$, number of LI columns $=2$ (i) $r=3$, nullity $=1$, number of LI rows $=3$, number of

LI columns $= 3$ **3.** (a) $r(\mathbf{A}) = 1$, $r(\mathbf{A}|\mathbf{c}) = 1$ so consistent; p-parameter family of solutions with $p = n - r = 4 - 1 = 3$
(e) $r(\mathbf{A}) = 2$, $r(\mathbf{A}|\mathbf{c}) = 3$ so inconsistent (i) $r(\mathbf{A}) = 3$, $r(\mathbf{A}|\mathbf{c}) = 3$ so consistent; p-parameter family of solutions with $p = n - r = 4 - 3 = 1$. **8.** (a) $r = 2$ so (Theorem 10.5.2) LD **16.** (a) 4; e.g., $H_2 + O_2 \rightleftharpoons 2\,OH$, $O_2 + 2H_2O \rightleftharpoons 4OH$, $H + 3OH \rightleftharpoons 2H_2O + O$, $OH \rightleftharpoons O + H$

Section 10.6

1. (a) $\dfrac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ (g) $\begin{bmatrix} 0 & 1/2 & -5/6 \\ 1 & 0 & 0 \\ 0 & 0 & 1/3 \end{bmatrix}$ (q) $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ **5.** (a) $x_1 = 24/13$, $x_2 = -6/13$ (d) $x_1 = 72/11$, $x_2 = -3/22$ **7.** (a) $\mathbf{A} = \begin{bmatrix} 2/3 & -1/3 \\ 1 & -1 \end{bmatrix}$ **11.** (a) $\begin{bmatrix} 1 & -2 & 13 \\ 0 & 1 & -8 \\ 0 & 0 & 1 \end{bmatrix}$

Section 10.7

1. $\mathbf{Q} = \begin{bmatrix} 1 & 2 \\ 1 & -1 \end{bmatrix}$. Not orthogonal because the column vectors are not ON. $[\mathbf{x}]_{B'} = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$, $[\mathbf{x}]_B = \begin{bmatrix} 8/3 \\ -1/3 \end{bmatrix}$
3. (a) $\mathbf{Q}= \begin{bmatrix} 2/\sqrt{5} & 1/\sqrt{5} \\ 1/\sqrt{5} & -2/\sqrt{5} \end{bmatrix}$, which is orthogonal. **4.** (b) $[\mathbf{x}]_{B'} = [\sqrt{2}, 2, 5/\sqrt{3}, -10/\sqrt{6}\,]^T$ **5.** (a) No
(e) Yes **9.** $\mathbf{Q}^n = \begin{bmatrix} \cos n\theta & \sin n\theta \\ -\sin n\theta & \cos n\theta \end{bmatrix}$

Section 10.8

2. (a) Nonlinear (d) Linear **3.** (b) The nth-order identity matrix **I** **5.** (a) dim $R = 2$, dim $K = 1$, dim $V = 3$. **F** is onto, is not one-to-one and is not invertible. A basis for K is $[0, -1, 1]^{\mathrm{T}}$, and a basis for R is $\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.
(d) dim $R = 2$, dim $K = 0$, dim $V = 2$. **F** is not onto, is one-to-one, and is not invertible. A basis for R is $[4, 3, 0]^{\mathrm{T}}$, $[0, 5, -4]^{\mathrm{T}}$, and K has no basis. **6.** (a) To be one-to-one we need $r(\mathbf{A}) = 2$. Also, to be onto we need $r(\mathbf{A}) = 2$. Thus, if $r(\mathbf{A}) = 2$ (i.e., **A** is any 2 × 2 matrix with nonzero determinant) then **F** is one-to-one and onto; if $r(\mathbf{A}) < 2$ (e.g., a 2 × 2 matrix with only one nonzero element) then it is neither. Can't be one and not the other. **7.** $\mathbf{A} = \begin{bmatrix} v_1^2 & v_1v_2 & v_1v_3 \\ v_2v_1 & v_2^2 & v_2v_3 \\ v_3v_1 & v_3v_2 & v_3^2 \end{bmatrix}$ **10.** (b) $\mathbf{A} = \begin{bmatrix} 2L_1^2 - 1 & 2L_2L_1 \\ 2L_1L_2 & 2L_2^2 - 1 \end{bmatrix}$ **11.** (a) $(\mathbf{GF})(\mathbf{x}) = \begin{bmatrix} -20 \\ -40 \end{bmatrix}$, **GF** transformation matrix $= \begin{bmatrix} 2 & 10 & 6 & -4 \\ 4 & 20 & 12 & -8 \end{bmatrix}$ **13.** (c) $\mathbf{F}(\mathbf{X}) = [2, 1, \sqrt{2} - 3, 1]^{\mathrm{T}}$
(d) $\mathbf{F}(\mathbf{X}_{\mathrm{p}}) = [1.2439,\ 3.5859,\ 2.2637,\ 1]^{\mathrm{T}}$, $\mathbf{F}(\mathbf{X}_e) = [1.5394,\ 3.7757,\ 1.3275,\ 1]^{\mathrm{T}}$

Chapter 11

Section 11.2

3. (a) $\lambda_1 = 0$ (multiplicity 2), eigenspace is $\mathbf{e}_1 = \alpha[1, 0]^{\mathrm{T}} + \beta[0, 1]^{\mathrm{T}}$ where α and β are arbitrary or, equivalently, span $\{[1, 0]^{\mathrm{T}}, [0, 1]^{\mathrm{T}}\}$. Basis for eigenspace is $\{[1, 0]^{\mathrm{T}}, [0, 1]^{\mathrm{T}}\}$ (i) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[0, -1, 1]^{\mathrm{T}}$ with basis $[0, -1, 1]^{\mathrm{T}}$; $\lambda_2 = 2$ (multiplicity 2), $\mathbf{e}_2 = \beta[0, 1, 1]^{\mathrm{T}} + \gamma[1, 0, 0]^{\mathrm{T}}$ with basis $\{[0, 1, 1]^{\mathrm{T}}, [1, 0, 0]^{\mathrm{T}}\}$. (n) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[13, 1, 0]^{\mathrm{T}}$ with basis $\{[13, 1, 0]^{\mathrm{T}}\}$, $\lambda_2 = 2$, $\mathbf{e}_2 = \beta[3, 1, 2]^{\mathrm{T}}$ with basis $\{[3, 1, 2]^{\mathrm{T}}\}$, $\lambda_3 = 5$, $\mathbf{e}_3 = \gamma[3, 1, 5]^{\mathrm{T}}$ with basis $\{[3, 1, 5]^{\mathrm{T}}\}$
5. (a) No (b) Yes, with $\lambda = -6$. **6.** (a) $\mathbf{e} = \alpha[1, 1, -1, -1]^{\mathrm{T}}$ **16.** (a) **A** has $\lambda_1 = 1$, $\mathbf{e}_1 = \alpha[1, -1, 0]^{\mathrm{T}}$; $\lambda_2 = 2$;

$\mathbf{e}_2 = \beta[0,1,-1]^{\mathrm{T}}$; $\lambda_3 = 3$, $\mathbf{e}_3 = \gamma[0,0,1]^{\mathrm{T}}$. Hence, $\mathbf{A}^{10}$ has $\lambda_1 = 1$, $\lambda_2 = 2^{10}$, $\lambda_3 = 3^{10}$, with the same eigenspaces as $\mathbf{A}$. **22.** (a) $\lambda = 1, (-1+\sqrt{3})/4, (-1-\sqrt{3})/4$ **23.** (a) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[2,-1]^{\mathrm{T}}$; $\lambda_2 = 7$, $\mathbf{e}_2 = \beta[1,3]^{\mathrm{T}}$ so $[x,y]^{\mathrm{T}} = \alpha[2,-1]^{\mathrm{T}}e^{0t} + \beta[1,3]Te^{7t}$ or $x(t) = 2\alpha + \beta e^{7t}$, $y(t) = -\alpha + 3\beta e^{7t}$. (c) $\lambda_1 = r^2 = 5$, $\mathbf{e}_1 = \alpha\begin{bmatrix} 1 \\ 3 \end{bmatrix}$ gives $\begin{bmatrix} x \\ y \end{bmatrix} = C_1\begin{bmatrix} 1 \\ 3 \end{bmatrix}e^{+\sqrt{5}t}$, $C_2\begin{bmatrix} 1 \\ 3 \end{bmatrix}e^{-\sqrt{5}t}$
$\lambda_2 = r^2 = -1$, $\mathbf{e}_2 = \beta\begin{bmatrix} 1 \\ -3 \end{bmatrix}$ gives $\begin{bmatrix} x \\ y \end{bmatrix} = C_3\begin{bmatrix} 1 \\ -3 \end{bmatrix}e^{it}$, $C_4\begin{bmatrix} 1 \\ -3 \end{bmatrix}e^{-it}$ so, by superposition, $\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = C_1\begin{bmatrix} 1 \\ 3 \end{bmatrix}e^{\sqrt{5}t} + C_2\begin{bmatrix} 1 \\ 3 \end{bmatrix}e^{-\sqrt{5}t} + C_3\begin{bmatrix} 1 \\ -3 \end{bmatrix}e^{it} + C_4\begin{bmatrix} 1 \\ -3 \end{bmatrix}e^{-it}$ or $x(t) = C_1e^{\sqrt{5}t} + C_2e^{-\sqrt{5}t} + C_3e^{it} + C_4e^{-it}$; $y(t) = 3C_1e^{\sqrt{5}t} + 3C_2e^{-\sqrt{5}t} - 3C_3e^{it} - 3C_4e^{-it}$ or $x(t) = C_5\cosh\sqrt{5}t + C_6\sinh\sqrt{5}t + C_7\cos t + C_8\sin t$; $y(t) = 3C_5\cosh\sqrt{5}t + 3C_6\sinh\sqrt{5}t - 3C_7\cos t - 3C_8\sin t$ (e) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[1,-4,5]^{\mathrm{T}}$; $\lambda_2 = 1$, $\mathbf{e}_2 = \beta[1,-3,2]^{\mathrm{T}}$; $\lambda_3 = 2$, $\mathbf{e}_3 = \gamma[1,-2,1]^{\mathrm{T}}$ so $[x(t),y(t),z(t)]^{\mathrm{T}} = \alpha[1,-4,5]^{\mathrm{T}} + \beta[1,-3,2]^{\mathrm{T}}e^t + \gamma[1,-2,1]^{\mathrm{T}}e^{2t}$ or $x(t) = \alpha + \beta e^t + \gamma e^{2t}$, $y(t) = -4\alpha - 3\beta e^t - 2\gamma e^{2t}$, $z(t) = 5\alpha + 2\beta e^t + \gamma e^{2t}$ **25.** (a) $\det(\mathbf{A} - \lambda\mathbf{B}) = (5-8\lambda)^2 - (1+4\lambda)^2 = 0$ so $\lambda = 1/3, 3/2$: $\lambda_1 = 1/3$, $\mathbf{e}_1 = \alpha[1,-1]^{\mathrm{T}}$; $\lambda_2 = 3/2$, $\mathbf{e}_2 = \beta[1,1]^{\mathrm{T}}$ **27.** (a) $\lambda = \pm 2$, saddle (d) $\lambda = -2, -4$, stable node

Section 11.3

1. (a) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[1,-1]^{\mathrm{T}}$; $\lambda_2 = 2$, $\mathbf{e}_2 = \beta[1,1]^{\mathrm{T}}$. Orthogonal basis: $\{[1,-1]^{\mathrm{T}},[1,1]^{\mathrm{T}}\}$ (d) $\lambda_1 = -2$, $\mathbf{e}_1 = \alpha[1,-1,0]^{\mathrm{T}} + \beta[0,0,1]^{\mathrm{T}}$; $\lambda_2 = 2$, $\mathbf{e}_2 = \gamma[1,1,0]^{\mathrm{T}}$. Orthogonal basis: $\{[1,-1,0]^{\mathrm{T}}, [0,0,1]^{\mathrm{T}}, [1,1,0]^{\mathrm{T}}\}$
(g) Orthogonal basis: $\{[1,0,0,1]^{\mathrm{T}}, [0,1,1,0]^{\mathrm{T}}, [1,0,0,-1]^{\mathrm{T}}, [0,1,-1,0]^{\mathrm{T}}\}$ **2.** (a) $\lambda_1 = 0$, $\mathbf{e}_1 = \alpha[1,-3]^{\mathrm{T}}$; $\lambda_2 = 1$, $\mathbf{e}_2 = \beta[0,1]^{\mathrm{T}}$. No. **7.** (a) $\alpha = 5/2, \phi_1 = \pi/2, \beta = -1/2, \phi_2 = \pi/2$ **8.** (b) $\lambda = \omega^2$. $\lambda_1 = 2-\sqrt{2} = 0.586$, $\omega_1 = 0.765$, $\mathbf{e}_1 = \alpha[1,\sqrt{2},1]^{\mathrm{T}}$; $\lambda_2 = 2$, $\omega_2 = 1.414$, $\mathbf{e}_2 = \beta[1,0,-1]^{\mathrm{T}}$; $\lambda_3 = 2+\sqrt{2} = 3.414$, $\omega_3 = 1.848$, $\mathbf{e}_3 = \gamma[1,-\sqrt{2},1]^{\mathrm{T}}$.
$x_1(t) = \alpha\sin(0.765t + \phi_1) + \beta\sin(1.414t + \phi_2) + \gamma\sin(1.848t + \phi_3)$
$x_2(t) = \sqrt{2}\,\alpha\sin(0.765t + \phi_1) + 0\beta\sin(1.414t + \phi_2) - \sqrt{2}\,\gamma\sin(1.848t + \phi_3)$
$x_3(t) = \alpha\sin(0.765t + \phi_1) - \beta\sin(1.414t + \phi_2) + \gamma\sin(1.848t + \phi_3)$
The natural frequencies are ω_1, ω_2, ω_3 (above) and their mode shapes are given by their corresponding eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$.
(c) Low mode: $x_1(0) = 1, x_2(0) = \sqrt{2}, x_3(0) = 1, x_1'(0) = x_2'(0) = x_3'(0) = 0$
Middle mode: $x_1(0) = 1, x_2(0) = 0, x_3(0) = -1, x_1'(0) = x_2'(0) = x_3'(0) = 0$
High mode: $x_1(0) = 1, x_2(0) = -\sqrt{2}, x_3(0) = 1, x_1'(0) = x_2'(0) = x_3'(0) = 0$
9. Partial answer: lowest natural frequency $= 0.518$, corresponding mode shape $= \alpha[1,\sqrt{3},2,\sqrt{3},1]^{\mathrm{T}}$.
11. (c) $[1,0,0]^{\mathrm{T}}$ is an eigenvector, corresponding to $\lambda = 2$, so the values of $R(\mathbf{x})$ for $\mathbf{x} = [0.4,0.3,0.3]^{\mathrm{T}}$, $[0.6,0.2,0.2]^{\mathrm{T}}$, $[0.8,0.1,0.1]^{\mathrm{T}}$, $[0.96,0.02,0.02]^{\mathrm{T}}$, $[1,0,0]^{T}$, should converge to 2. In fact, they are -0.65, 1.09, 1.85, 1.996, 2. Similarly, $[0,1,1]^{\mathrm{T}}$ is an eigenvector corresponding to $\lambda = -3$, so the values of $R(\mathbf{x})$, for $\mathbf{x} = [0,1,1.4]^{\mathrm{T}}, [0,1,1.1]^{\mathrm{T}}, [0,1,1]^{\mathrm{T}}$, should converge to -3. In fact, they are -3.08, -3.04, -3. **13.** (a) $\mathbf{x}^{(0)} = [1,0,0]^{\mathrm{T}}$, $\mathbf{x}^{(1)} = [2,1,-1]^{\mathrm{T}}$, $\mathbf{x}^{(2)} = [6,3,-3]^{\mathrm{T}}$ which is an exact multiple of $\mathbf{x}^{(1)}$, namely, 3. Thus, one eigenpair is $\lambda = 3$, $\mathbf{e} = [2,1,-1]^{\mathrm{T}}$. Repeat with $\mathbf{x}^{(0)} = [0,1,0]^{\mathrm{T}}$. Then $\mathbf{x}^{(1)} = [1,4,3]^{\mathrm{T}}$, $\mathbf{x}^{(2)} = [3,26,23]^{\mathrm{T}}$, $\mathbf{x}^{(3)} = [9,176,167]^{\mathrm{T}}$, $\mathbf{x}^{(4)} = [27,1214,1187]^{\mathrm{T}}$ so $\lambda \approx \mathbf{x}^{(3)T}\mathbf{A}\mathbf{x}^{(3)}/\mathbf{x}^{(3)T}\mathbf{x}^{(3)} = \mathbf{x}^{(3)T}\mathbf{x}^{(4)}/\mathbf{x}^{(3)T}\mathbf{x}^{(3)} = 412136/58946 = 6.992$ and $\mathbf{e} \approx [27,1214,1187]^{\mathrm{T}}$. Finally, repeating with $\mathbf{x}^{(0)} = [0,0,1]^{\mathrm{T}}$ gives convergence to the same eigenpair as that obtained from $\mathbf{x}^{(0)} = [0,1,0]^{\mathrm{T}}$. Since $\mathbf{x}^{(0)} = [1,0,0]^{\mathrm{T}}$ gave convergence to the subdominate eigenvalue $\lambda = 3$, $[1,0,0]^{\mathrm{T}}$ must be orthogonal to the eigenvector corresponding to the dominate eigenvalue. Indeed it is, because the exact values are $\lambda_1 = 7$, $\mathbf{e}_1 = [0,1,1]^{\mathrm{T}}$; $\lambda_2 = 3$, $\mathbf{e}_2 = [2,1,-1]$; $\lambda_3 = 0$, $\mathbf{e}_3 = [1,-1,1]^{\mathrm{T}}$.

14. (a) $\mathbf{A}^4 = \begin{bmatrix} 54 & 27 & -27 \\ 27 & 1214 & 1187 \\ -27 & 1187 & 1214 \end{bmatrix}$, $\mathbf{x}^{(0)} = [0,1,0]^{\mathrm{T}}$, $\mathbf{x}^{(1)} = [27,1214,1187]^{\mathrm{T}}$,

$\mathbf{x}^{(2)} = [2187, 2883493, 2881307]^{\mathrm{T}}$, $\lambda \approx \mathbf{x}^{(1)T}\mathbf{x}^{(2)}/\mathbf{x}^{(1)T}\mathbf{x}^{(1)} = 2400.12$. But this is the fourth power of $\mathbf{A}$'s λ, so $\lambda = (2400.12)^{1/4} = 6.9994$, with $\mathbf{e} \approx [2187, 2883493, 2881307]^{\mathrm{T}}$. Convergence is more rapid using $\mathbf{A}^4$ because its eigenvalues are $7^4 = 2401$, $3^4 = 81$, $0^4 = 0$, so the largest eigenvalue of $\mathbf{A}^4$ is more dominant than the largest eigenvalue of $\mathbf{A}$; i.e., $(7/3)^4 \gg 7/3$. **15.** (a) For $\mathbf{A}$, $\lambda_1 = 7$, $\mathbf{e}_1 = [0,1,1]^{\mathrm{T}}$; $\lambda_2 = 3$, $\mathbf{e}_2 = [2,1,-1]^{\mathrm{T}}$; $\lambda_3 = 0$, $\mathbf{e}_3 = [1,-1,1]^{\mathrm{T}}$. Thus, $\mathbf{c} = [1,2,3]^{\mathrm{T}} = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3$ where (because the $\mathbf{e}_j$'s are orthogonal) $c_1 = \mathbf{c}\cdot\mathbf{e}_1/\mathbf{e}_1\cdot\mathbf{e}_1 = 5/2$, $c_2 = 1/6$, $c_3 = 2/3$. Then $a_j = c_j/(\lambda_j - \Lambda)$ gives $a_1 = 1/2$, $a_2 = 1/6$, $a_3 = -1/3$, hence the unique solution $\mathbf{x} = \frac{1}{2}\mathbf{e}_1 + \frac{1}{6}\mathbf{e}_2 - \frac{1}{3}\mathbf{e}_3 = [0,1,0]^{\mathrm{T}}$. (b) This time $\Lambda = 3$ coincides with λ_2. $\mathbf{c}\cdot\mathbf{e}_2 = 6 \neq 0$. Hence, no solution. (c) Again $\Lambda = 3$ coincides with λ_2, but this time $\mathbf{c}\cdot\mathbf{e}_2 = 0$. Hence, nonunique solution. $c_1 = 2$, $c_2 = 0$, $c_3 = 1$ so $a_1 = c_1/(\lambda_1 - \Lambda) = 1/2$, $a_2 = \alpha$ (arbitrary), $a_3 = c_3/(\lambda_3 - \Lambda) = -1/3$ and $\mathbf{x} = \frac{1}{2}\mathbf{e}_1 + \alpha\,\mathbf{e}_2 - \frac{1}{3}\mathbf{e}_3 = [2\alpha - 1/3, \alpha + 5/6, -\alpha + 1/6]^{\mathrm{T}}$. **17.** (b) $\lambda_1 = 0$, $\mathbf{e}_1 = [1,-1]^{\mathrm{T}}$; $\lambda_2 = 5/6$, $\mathbf{e}_2 = [2,3]^{\mathrm{T}}$. $\mathbf{e}_1\cdot\mathbf{M}\mathbf{e}_2 = [1,-1]^{\mathrm{T}}\cdot[6,6]^{\mathrm{T}} = 0$ and $\mathbf{e}_2\cdot\mathbf{M}\mathbf{e}_1 = 0$ too.

Section 11.4

1. (a) $\lambda_1 = 2$, $\mathbf{e}_1 = [1,0]^{\mathrm{T}}$; $\lambda_2 = 0$, $\mathbf{e}_2 = [3,2]^{\mathrm{T}}$. Thus, $\mathbf{Q} = \begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$ (d) $\lambda_1 = 0$, $\mathbf{e}_1 = [1,-1,0]^{\mathrm{T}}$; $\lambda_2 = -1$, $\mathbf{e}_2 = [1,1,-1]^{\mathrm{T}}$; $\lambda_3 = 2$, $\mathbf{e}_3 = [1,1,2]^{\mathrm{T}}$. Can use these $\mathbf{e}_j$'s as columns of $\mathbf{Q}$ but since we are asked to compute $\mathbf{Q}^{-1}$ it is best to use the *normalized* $\mathbf{e}_j$'s instead, so that $\mathbf{Q}^{-1}$ is simply $\mathbf{Q}^T$ (since the $\mathbf{e}_j$'s are ON). Then $\mathbf{Q} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ -1/\sqrt{2} & 1/\sqrt{3} & 1/\sqrt{6} \\ 0 & -1/\sqrt{3} & 2/\sqrt{6} \end{bmatrix}$ and $\mathbf{D} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ **2.** (a) $x(t) = Ae^{2t} + B$, $y(t) = Ae^{2t} - B$

5. (a) The ODE's are $x'' + \frac{1}{4}x - \frac{\sqrt{3}}{4}y = 0$ and $y'' - \frac{\sqrt{3}}{4}x + \frac{39}{4}y = 0$. $\omega_1 = \sqrt{\lambda_1} = \sqrt{0.2303} = 0.4799$, $\mathbf{e}_1 = [21.98, 1]^{\mathrm{T}}$; $\omega_2 = \sqrt{\lambda_2} = \sqrt{9.770} = 3.126$, $\mathbf{e}_2 = [0.04548, -1]^{\mathrm{T}}$. **8.** (a) $\lambda_1 = 4$, $\mathbf{e}_1 = [1,1,1]^{\mathrm{T}}$; $\lambda_2 = \lambda_3 = -2$, $\mathbf{e} = \alpha[1,-1,0]^{\mathrm{T}} + \beta[1,0,-1]^{\mathrm{T}}$, from which we can form the orthogonal pair $\mathbf{e}_2 = [1,-1,0]^{\mathrm{T}}$ and $\mathbf{e}_3 = [1,1,-2]^{\mathrm{T}}$. Normalize these so $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}$. Then

$$\mathbf{A}^{1000} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & -1/\sqrt{2} & 1/\sqrt{6} \\ 1/\sqrt{3} & 0 & -2/\sqrt{6} \end{bmatrix} \begin{bmatrix} 4^{1000} & 0 & 0 \\ 0 & (-2)^{1000} & 0 \\ 0 & 0 & (-2)^{1000} \end{bmatrix} \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} (4^{1000} + 2^{1001}) & (4^{1000} - 2^{1000}) & (4^{1000} - 2^{1000}) \\ (4^{1000} - 2^{1000}) & (4^{1000} + 2^{1001}) & (4^{1000} - 2^{1000}) \\ (4^{1000} - 2^{1000}) & (4^{1000} - 2^{1000}) & (4^{1000} + 2^{1001}) \end{bmatrix}$$

9. (a) $I_{xx} = \sigma$, $I_{xy} = 9\sigma/4$, $I_{yy} = 9\sigma$, $I_{zz} = 10\sigma$, $I_{xz} = I_{yz} = 0$ so $I = \sigma\begin{bmatrix} 1 & -9/4 & 0 \\ -9/4 & 9 & 0 \\ 0 & 0 & 10 \end{bmatrix}$. $\lambda_1 = 9.589\sigma$, $\mathbf{e}_1 = [1, -3.818, 0]^{\mathrm{T}}$; $\lambda_2 = 0.411\sigma$, $\mathbf{e}_2 = [1, 0.262, 0]^{\mathrm{T}}$; $\lambda_3 = 10\sigma$, $\mathbf{e}_3 = [0,0,1]^{\mathrm{T}}$ so $I_{x'x'} = 9.589\sigma$, $I_{y'y'} = 0.411\sigma$, $I_{z'z'} = 10\sigma$ where $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, $\hat{\mathbf{e}}_3$ are unit vectors in the x', y', z' coordinate directions.

Section 11.5

1. (a) $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{f}(t)\begin{bmatrix} 0 \\ e^{3t} \end{bmatrix}$, $\mathbf{c} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Then $\lambda_1 = 2$, $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; $\lambda_2 = 0$, $\mathbf{e}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$, $\mathbf{Q} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{Q}^{-1} = \begin{bmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{bmatrix}$, $\mathbf{x}(t) = \mathbf{0} + \int_0^t \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} e^{2(t-\tau)} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 0 \\ e^{3\tau} \end{bmatrix} d\tau$ gives $x(t) = \frac{1}{3}e^{3t} - \frac{1}{2}e^{2t} + \frac{1}{6}$, $y(t) = \frac{2}{3}e^{3t} - \frac{1}{2}e^{2t} - \frac{1}{6}$. (i) Let $x_1(t) \equiv x(t)$, $x_2(t) \equiv x'(t)$. Then $x(t) = x_1(t) = \frac{1}{6}e^{t} - \frac{1}{2}e^{-t} + \frac{1}{3}e^{-2t}$ [and $x'(t) = x_2(t) = \frac{1}{6}e^{t} + \frac{1}{2}e^{-t} - \frac{2}{3}e^{-2t}$]. **3.** (a) $\lambda_1 = 1$, $\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$; $\lambda_2 = 3$, $\mathbf{e}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; $\mathbf{Q} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\mathbf{Q}^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$,

$\mathbf{D} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$, $e^{\mathbf{A}} = \mathbf{Q}e^{\mathbf{D}}\mathbf{Q}^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} e & 0 \\ 0 & e^3 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e & e^3 - e \\ 0 & e^3 \end{bmatrix}$.
(d) $e^{\mathbf{A}} = \dfrac{e}{2}\begin{bmatrix} e^2+1 & e^2-1 \\ e^2-1 & e^2+1 \end{bmatrix}$.

Section 11.6

1. (a) $\mathbf{A} = \begin{bmatrix} 2 & 1/2 \\ 1/2 & 4 \end{bmatrix}$ (d) $\mathbf{A} = \begin{bmatrix} 1 & 3/2 & 0 \\ 3/2 & 0 & 0 \\ 0 & 0 & -4 \end{bmatrix}$ **2.** (a) $2x_1^2 + 4x_2^2 + x_1x_2 = \lambda_1\tilde{x}_1^2 + \lambda_2\tilde{x}_2^2 = \frac{3+\sqrt{5}}{2}\tilde{x}_1^2 + \frac{3-\sqrt{5}}{2}\tilde{x}_2^2$ where $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ and $\mathbf{Q} = \begin{bmatrix} a & b \\ (2+\sqrt{5})a & (2-\sqrt{5})b \end{bmatrix}$ where $a = 1/\sqrt{10+4\sqrt{5}}$ and $b = 1/\sqrt{10-4\sqrt{5}}$. $\lambda_1 > 0$, $\lambda_2 > 0$ so positive definite. (g) $3x_1x_2 = \frac{3}{2}\tilde{x}_1^2 - \frac{3}{2}\tilde{x}_2^2$ where $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ and $\mathbf{Q} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$. Neither positive nor negative definite. **4.** (a) $3x^2 + 2y^2 - 2xy = \frac{5+\sqrt{5}}{2}\tilde{x}^2 + \frac{5-\sqrt{5}}{2}\tilde{y}^2 = 6$ is an ellipse with intercepts at $\pm\sqrt{12/[5+\sqrt{5}\,]}$ on $\tilde{x}$ axis and $\pm\sqrt{12/[5-\sqrt{5}\,]}$ on $\tilde{y}$ axis. The positive $\tilde{x}$ and $\tilde{y}$ axes are oriented in the x, y plane in the directions of the orthogonal eigenvectors $\mathbf{e}_1 = [2, 1-\sqrt{5}\,]^{\mathrm{T}}$ and $\mathbf{e}_2 = [2, 1+\sqrt{5}\,]^{\mathrm{T}}$.
(d) $x^2 + y^2 - 10xy = 6\tilde{x}^2 - 4\tilde{y}^2 = 4$ is an hyperbola with intercepts at $\pm 2/\sqrt{6}$ on $\tilde{y}$ axis. The positive $\tilde{x}$ and $\tilde{y}$ axes are oriented in the x, y plane in the directions of the orthogonal eigenvectors $\mathbf{e}_1 = [1, -1]^{\mathrm{T}}$ and $\mathbf{e} = [1, 1]^{\mathrm{T}}$.

Chapter 12

Section 12.2

1. (a) $\hat{\mathbf{u}} = \left[\frac{1}{\sqrt{2}}, \frac{i}{\sqrt{2}}\right]$ **2.** (a) The set is a basis for $\mathbb{C}^3$. **4.** (a) $\mathbf{u} = \frac{4-5i}{2}\mathbf{e}_1 + \frac{8i}{9}\mathbf{e}_2 - \frac{9-4i}{18}\mathbf{e}_3$ **5.** $(\alpha\mathbf{x}\cdot\mathbf{y}) = \sum_1^n \alpha x_j\bar{y}_j = \alpha\sum_1^n x_j\bar{y}_j = \alpha(\mathbf{x}\cdot\mathbf{y})$, $\mathbf{x}\cdot(\alpha\mathbf{y}) = \sum_1^n x_j\overline{\alpha y}_j = \bar{\alpha}\sum_1^n x_j\bar{y}_j = \bar{\alpha}(\mathbf{x}\cdot\mathbf{y})$

Section 12.3

1. (a) $\mathbf{A}^{-1} = \dfrac{8+i}{65}\begin{bmatrix} 4 & -i \\ -1 & 2 \end{bmatrix}$ **3.** (a) $\lambda_1 = 3 + \sqrt[4]{2}e^{\pi i/8}$, $\mathbf{e}_1 = \alpha\left[1, \sqrt[4]{2}\sin\frac{\pi}{8} - (1 + \sqrt[4]{2}\cos\frac{\pi}{8})i\right]^{\mathrm{T}}$, $\lambda_2 = 3 - \sqrt[4]{2}e^{\pi i/8}$, $\mathbf{e}_2 = \beta\left[1, -\sqrt[4]{2}\sin\frac{\pi}{8} + (-1 + \sqrt[4]{2}\cos\frac{\pi}{8})i\right]^{\mathrm{T}}$ **5.** (b) $\lambda_1 = 2+i$, $\mathbf{e}_2 = \alpha\begin{bmatrix} 1 \\ i \end{bmatrix}$; $\lambda_2 = 2-i$, $\mathbf{e}_1 = \beta\begin{bmatrix} 1 \\ -i \end{bmatrix}$ **6.** $(a-d)^2 + 4bc < 0$ **7.** (a) $\mathbf{A} = \begin{bmatrix} 2 & 1+i \\ 1-i & 3 \end{bmatrix}$, $\lambda_1 = 1$, $\mathbf{e}_1 = \alpha\begin{bmatrix} 1+i \\ -1 \end{bmatrix}$, $\lambda_2 = 4$, $\mathbf{e}_2 = \beta\begin{bmatrix} 1 \\ 1-i \end{bmatrix}$ so $F = |x_1'|^2 + 4|x_2'|^2$, where $\mathbf{x} = \mathbf{U}\mathbf{x}' = \frac{1}{\sqrt{3}}\begin{bmatrix} 1+i & 1 \\ -1 & 1-i \end{bmatrix}\mathbf{x}'$, and $\mathbf{x}' = \mathbf{U}^{-1}\mathbf{x} = \mathbf{U}^*\mathbf{x} = \frac{1}{\sqrt{3}}\begin{bmatrix} 1-i & -1 \\ 1 & 1+i \end{bmatrix}\mathbf{x}$ **8.** (b) $\mathbf{A}^{1000} = \dfrac{1}{5}\begin{bmatrix} 4+6^{1000} & 2i(1-6^{1000}) \\ 2i(6^{1000}-1) & 4(6^{1000})+1 \end{bmatrix}$ **12.** Neither **15.** (a) $c_2 = c_1$ (b) None

Chapter 13

Section 13.2

1. (a) $\sqrt{21}$ (d) $5\sqrt{2}$ **2.** Yes, because $d(P, P_0) = 0.36 < r$ **6.** (a) Connected, neither (f) Not connected, closed (k) Connected, open (n) Connected, closed **12.** (a) The points $x = 0, 1, -2$ (c) The lines $x + y = n\pi/2$ $(n = \pm 1, \pm 3, \ldots)$

Section 13.3

1. (a) $f_x = 3x^2y^5$, $f_y = 5x^3y^4$, $f_{xy} = f_{yx} = 15x^2y^4$, $f_{xx} = 6xy^5$, $f_{yy} = 20x^3y$ **4.** $\alpha = 1 - n/2$: for $n = 2$, $\alpha = 0$; for $n = 3$, $\alpha = -1/2$; etc.

Section 13.4

1. $F'(t) = [4x^3 \cos(x^4 + 3y)](5) + [3\cos(x^4 + 3y)](2t) = (2500t^3 + 6t)\cos(625t^4 + 3t^2 + 3)$ **2.** (a) $F'(t) = (ye^{xy})(2t) + (xe^{xy})(3\cos 3t)$ **3.** (a) $G'(s) = (\sin s \cos s + 3\sin 6s)/\sqrt{\sin^2 s - \cos 6s}$ **4.** (a) $R'(2) = 0.3553$

Section 13.5

1. (a) TS $e^{-2x}\big|_{x=0} = 1 - 2x + 2x^2 - \frac{4}{3}x^3 + \cdots$ (d) TS $\ln x\big|_{x=2} = \ln 2 + \frac{1}{2}(x-2)^2 - \frac{1}{8}(x-2)^2 + \frac{1}{24}(x-2)^3 - \cdots$ **2.** (a) $1 + x^5 + x^{10} + x^{15} + \cdots$ **9.** (a) $0 = f(x, y, z) = x^3yz + 6 = 0 - 18(x-1) - 2(y-3) + 3(z-2) + \cdots$ gives the tangent plane $18x + 2y - 3z = 30$. **11.** (a) First, expand in x, say: $x^5y^4 - y = [y^4 - y] + [5y^4](x - 1) + \frac{1}{2!}[20y^4](x-1)^2 + \frac{1}{3!}[60y^4](x-1)^3 + \cdots$. Next, expand each square-bracketed coefficient in y: $x^5y^4 - y = [14 + 31(y-2) + 24(y-2)^2 + 8(y-2)^3 + \cdots] + [80 + 160(y-2) + 120(y-2)^2 + \cdots](x-1) + \frac{1}{2}[320 + 640(y - 2) + \cdots](x-1)^2 + \frac{1}{6}[960 + \cdots](x-1)^3 + \cdots$. Arrange terms in ascending order:

$x^5y^4 - y = 14 + 80(x-1) + 31(y-2) + 160(x-1)^2 + 160(x-1)(y-2) + 24(y-2)^2 + 160(x-1)^3 + 320(x - 1)^2(y-2) + 120(x-1)(y-2)^2 + 8(y-2)^3 + \cdots$. NOTE: Clearly, this series converges for all x, y, because it contains only a finite number of terms (since all x derivatives of $x^5y^4 - y$ above fifth–order and all y derivatives above fourth–order are zero); i.e., the series *terminates*.

Section 13.6

1. (a) f is C^1 in the entire plane, but $f_y(2, 1) = 0$ so the conditions are not met. Indeed, the graph of $x^2 + (y-1)^2 = 1$ is a circle of radius 1 centered at $(0, 1)$, with vertical tangent at $(1, 1)$. NOTE: Although the relation does not imply an implicit function $y(x)$ through $(1, 1)$ it does imply an implicit function $x(y)$ through that point.
(d) $y(x) = 1 + \frac{3}{2}(x+2) + \frac{48-9\pi^2}{32}(x+2)^2 + \cdots$ **2.** (a) $y' = y/(3y^2 - x)$, $y'' = -2xy/(3y^2 - x)^3$
3. (a) $y_x = -[y + \cos(x+z)]/x$, $z_x = -[y + \cos(x+z)]/[2z - \cos(x+z)]$ **4.** (a) Conditions not met because $\frac{\partial(f,g)}{\partial(u,v)} = 0$ at $(0, 0, 0, 0)$. NOTE: The conditions are stated to be sufficient, not necessary. In this example the conditions are not met so the theorm gives no information. **5.** (a) $\partial/\partial y$ gives $-1 + 2uu_y + 2vv_y = 0$ and $1 + 3u^2e^vu_y + u^3e^vv_y = 0$ so $u_y = (u^3 + 2ve^{-v})/(2u^4 - 6u^2v)$. **6.** (a) $\frac{\partial(f,g)}{\partial(u,v)} = \begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix} = \begin{vmatrix} 3v^2 & 6uv \\ 2u & -2v \end{vmatrix} = -6v^3 - 12u^2v$
12. (b) They are true. Since there are only the three independent variables p, T, v in (12.1), both $\frac{\partial v}{\partial p}$ and $\frac{\partial p}{\partial v}$ are computed with T fixed and are therefore numerical inverses of each other. Or, working them out, $\partial/\partial p$ and $\partial/\partial v$ of (12.1) give $f_p + f_vv_p = 0$ and $f_v + f_pp_v = 0$. Thus, $v_p = -f_p/f_v$ and $p_v = -f_v/f_p = 1/v_p$.
14. (a) $\frac{\partial}{\partial x} = \frac{1}{4}\frac{\partial}{\partial u} + \frac{1}{2}\frac{\partial}{\partial v}$ and $\frac{\partial}{\partial y} = \frac{1}{4}\frac{\partial}{\partial u} - \frac{1}{2}\frac{\partial}{\partial v}$ so $T_{xx} + T_{yy} = \left(\frac{1}{4}\frac{\partial}{\partial u} + \frac{1}{2}\frac{\partial}{\partial v}\right)\left(\frac{1}{4}T_u + \frac{1}{2}T_v\right) + \left(\frac{1}{4}\frac{\partial}{\partial u} - \frac{1}{2}\frac{\partial}{\partial v}\right)\left(\frac{1}{4}T_u - \frac{1}{2}T_v\right) = \frac{1}{8}T_{uu} + \frac{1}{2}T_{vv} = 0$ or $T_{uu} + 4T_{vv} = 0$ (e) $T_{uu} + \frac{1}{u}T_u + \frac{4}{u^2}T_{vv} = 0$

Section 13.7

1. (a) $f(x) = 0 + 0(x-1) +)(x-1)^2 + \frac{6}{3!}(x-1)^3 + \cdots$ so horizontal inflection point (c) maximum (e) minimum
2. (a) $x = 0$: $f(x) = 1 - x^3 + \cdots$ so horizontal inflection point (c) $x = 2$: $f(x) = e^{16} - 6e^{16}(x-2)^2 + \cdots$ so maximum;
$x = -2$: $f(x) = e^{-16} - 6e^{-16}(x-2)^2 + \cdots$ so minimum **5.** (a) $f_x = 4x + y = 0$ and $f_y = x + 2y - 7 = 0$ gives $x = 1$, $y = -4$. At $(1, -4)$, $f_{xx} = 4$, $f_{xy} = 1$, $f_{yy} = 2$ so

$\mathbf{A} = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$ with $\lambda = -1, 7$. Mixed signs, hence saddle. (g) $f_x = (2x + y)/(x^2 + xy + y^2 + 4) = 0$ and $f_y = (x + 2y)/(x^2 + xy + y^2 + 4) = 0$ give $x = y = 0$. At $(0, 0)$, $f_{xx} = 1/2$, $f_{xy} = 1/4$, $f_{yy} = 1/2$

so $\mathbf{A} = \begin{pmatrix} 1/2 & 1/4 \\ 1/4 & 1/2 \end{pmatrix}$ with $\lambda = 1/4, 3/4$. Both positive, hence minimum. **6.** (a) f_x, f_y, f_z are 0 at (0,0,0); $\lambda = 2, 1 \pm \sqrt{10}$ are of mixed sign, hence saddle. (d) f_x, f_y, f_z are 0 at (0,0,0); $\lambda =1, 1, 1$, hence minimum. **8.** $|\alpha| \leq 10$ **13.** (a) $x = 2/5, y = 4/5$ **16.** (a) $x = 13/5, y = -6/5$ **17.** (a) $x = 3, y = 1, z = 0$ **18.** $x = (\sqrt{7}-1)/3$, $y = [(\sqrt{7}-1)/3]^{3/2}$ **19.** max. at $(1/\sqrt{2},\ 0,\ 1/\sqrt{2})$, min. at $(-1/\sqrt{2},\ 0,\ -1/\sqrt{2})$

Section 13.8

1. (a) $\int_0^{t^2} x^2 \cos(tx^2)dx + 2t\sin(t^5)$ (b) $\int_{-2\alpha^2}^{-\alpha} x^3 e^{\alpha x^3}dx + (-1)e^{\alpha(-\alpha)^3} - (-4\alpha)e^{\alpha(-2\alpha^2)^3}$ **2.** (a) $0 + x + 0x^2 + \cdots$ (d) $1 + \frac{5}{3}x - \frac{19}{10}x^2 + \cdots$ **9.** We can evaluate $I(t)$ by using the change of variables $\sqrt{x} = u$ $(x = u^2)$, and find that $I(t) = 2e^{\sqrt{t}} - 2$. Thus, $I'(t) = e^{\sqrt{t}}/\sqrt{t}$. On the other hand, the Leibniz rule gives $I'(t) = e^{\sqrt{t}}/\sqrt{t}$, which is the same result. **11.** (a) Liebniz gives $u_{xx} = \frac{1}{4\sqrt{\pi\alpha^3 t^{3/2}}}\int_0^1 f(\xi)\left[-1 + \frac{(x-\xi)^2}{2\alpha^2 t}\right]e^{-(x-\xi)^2/(4\alpha^2 t)}d\xi$, $u_t = -\frac{1}{4\alpha\sqrt{\pi t^{3/2}}}\int_0^1 f(\xi)e^{-(x-\xi)^2/(4\alpha^2 t)}d\xi + \int_0^1 \frac{f(\xi)}{2\alpha\sqrt{\pi t}}\frac{(x-\xi)^2}{4\alpha^2 t^2}e^{-(x-\xi)^2/(4\alpha^2 t)}d\xi$ so $\alpha^2 u_{xx}$ does not equal u_t.

Chapter 14

Section 14.3

1. (a) $2\mathbf{u} - \mathbf{v} = 3\hat{\mathbf{i}} - 3\hat{\mathbf{j}} - 5\hat{\mathbf{k}}$, $\mathbf{u}\cdot\mathbf{v} = 4$, $\mathbf{v}\cdot\mathbf{u} = 4$, $\mathbf{u}\times\mathbf{v} = 4\hat{\mathbf{i}} - \hat{\mathbf{j}} + 3\hat{\mathbf{k}}$, $\mathbf{v}\times\mathbf{u} = -4\hat{\mathbf{i}} + \hat{\mathbf{j}} - 3\hat{\mathbf{k}}$, $\|\mathbf{u}\times\mathbf{v}\| = \sqrt{26}$ (h) $\mathbf{u}\times\mathbf{v} = 4\hat{\mathbf{i}} - \hat{\mathbf{j}} + 3\hat{\mathbf{k}}$ is perpendicular to $\mathbf{u}$ and $\mathbf{v}$. Check: $\mathbf{u}\cdot(\mathbf{u}\times\mathbf{v}) = 8 + 1 - 9 = 0$, $\mathbf{v}\cdot(\mathbf{u}\times\mathbf{v}) = 4 - 1 - 3 = 0$. (k) Area $= \sqrt{122}/2$, $\alpha(-8\hat{\mathbf{i}} - 7\hat{\mathbf{j}} + 3\hat{\mathbf{k}})$ for any $\alpha \neq 0$ **2.** (a) $\mathbf{M} = -3\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + \hat{\mathbf{k}}$ (d) $\mathbf{M} = -4\hat{\mathbf{i}} - \hat{\mathbf{j}}$ **3.** (a) No (c) Yes (e) No **4.** (a) $\sqrt{161}$ (d) $\sqrt{2873}$ **5.** (a) $\pm(2\hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}})/\sqrt{6}$ **8.** $\mathbf{x}\cdot\mathbf{w} = \{a[1,1,2] + b[3,2,-1]\}\cdot[2,4,3] = 0$ gives $12a + 11b = 0$ so $b = -12a/11$. Let $a = 11$; then $b = -12$ and $\mathbf{x} = [-25, -13, 34]$ (times an arbitrary nonzero scale factor)

Section 14.4

1. (a) $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w} = \mathbf{u}\times\mathbf{v}\cdot\mathbf{w} = -6$ **3.** (a) LHS $= \hat{\mathbf{i}}\times(\hat{\mathbf{j}}\times\hat{\mathbf{k}}) = \hat{\mathbf{i}}\times\hat{\mathbf{i}} = \mathbf{0}$; RHS $= (\hat{\mathbf{i}}\cdot\hat{\mathbf{k}})\hat{\mathbf{j}} - (\hat{\mathbf{i}}\cdot\hat{\mathbf{j}})\hat{\mathbf{k}} = \mathbf{0}$ **4.** (a) 3 (d) 30 **12.** (a) No (d) Yes

Section 14.5

1. (a) $\mathbf{u} = 2\tau\mathbf{B}$, $\mathbf{u}'' = 2\mathbf{B}$, $\|\mathbf{u}''\| = 2\|\mathbf{B}\|$ (d) $\mathbf{u}' = -\sin\tau\,\hat{\mathbf{i}} + \cos\tau\,\hat{\mathbf{j}} + \cos\tau\,\hat{\mathbf{k}}$, $\mathbf{u}'' = -\cos\tau\,\hat{\mathbf{i}} - \sin\tau\,\hat{\mathbf{j}} - \sin\tau\,\hat{\mathbf{k}}$, $\|\mathbf{u}''\| = \sqrt{1 + \sin^2\tau}$ **2.** (a) $(\mathbf{u}\cdot\mathbf{v})' = (2\tau^2\cos\tau + 3\tau^2)' = 4\tau\cos\tau - 2\tau^2\sin\tau + 6\tau$, $\mathbf{u}\cdot\mathbf{v}' + \mathbf{u}'\cdot\mathbf{v} = -2\tau^2\sin\tau + 3\tau + 4\tau\cos\tau + 3\tau$ (checks) (b) $(\mathbf{u}\times\mathbf{v})' = 2\tau\,\hat{\mathbf{i}} + (3\tau^2 - 6\cos\tau + 6\tau\sin\tau)\,\hat{\mathbf{j}} + (2\cos\tau - 2\tau\sin\tau)\,\hat{\mathbf{k}}$, $\mathbf{u}\times\mathbf{v}' + \mathbf{u}'\times\mathbf{v} = \tau\,\hat{\mathbf{i}} + (\tau^2 + 6\tau\sin\tau)\,\hat{\mathbf{j}} - 2\tau\sin\tau\,\hat{\mathbf{k}} + \tau\,\hat{\mathbf{i}} + (2\tau^2 - 6\cos\tau)\,\hat{\mathbf{j}} + 2\cos\tau\,\hat{\mathbf{k}}$ (checks)
7. (a) $(\mathbf{u}\cdot\mathbf{v})'' = (\mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}')' = (\mathbf{u}'\cdot\mathbf{v})' + (\mathbf{u}\cdot\mathbf{v}')' = \mathbf{u}''\cdot\mathbf{v} + \mathbf{u}'\cdot\mathbf{v}' + \mathbf{u}'\cdot\mathbf{v}' + \mathbf{u}\cdot\mathbf{v}'' = \mathbf{u}''\cdot\mathbf{v} + 2\mathbf{u}'\cdot\mathbf{v}' + \mathbf{u}\cdot\mathbf{v}''$

Section 14.6

1. (a) $r = 2\sqrt{2}$, $\theta = \tan^{-1}(2/2) = \pi/4$ and $5\pi/4$, $z = 3$. Discard $5\pi/4$ so $\theta = \pi/4$ rad $= 45°$.
(d) $r = 0$, θ not defined, $z = 5$ (g) $r = \sqrt{26}$, $\theta = \tan^{-1}(5/-1) = 1.768$ rad $= 101.3°$ **2.** (a) $\rho = 2$, $\phi = \pi/2$ rad $= 90°$, $\theta = 0$ (d) $\rho = \sqrt{14}$, $\phi = \cos^{-1}(3/\sqrt{14}) = 0.641$ rad $= 36.70°$, $\theta = \tan^{-1}(2/1) = 1.107$ rad $= 63.43°$ **3.** (a) Since $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ is an ON basis for 3- space we can write $\hat{\mathbf{e}}_\rho = (\hat{\mathbf{e}}_\rho\cdot\hat{\mathbf{i}})\,\hat{\mathbf{i}} + (\hat{\mathbf{e}}_\rho\cdot\hat{\mathbf{j}})\,\hat{\mathbf{j}} + (\hat{\mathbf{e}}_\rho\cdot\hat{\mathbf{k}})\,\hat{\mathbf{k}}$ and similarly for $\hat{\mathbf{e}}_\phi$ and $\hat{\mathbf{e}}_\theta$. Then $\hat{\mathbf{e}}\cdot\hat{\mathbf{i}} = (1)(1)\cos(\hat{\mathbf{e}}_\rho, \hat{\mathbf{i}}) = x/\rho = \sin\phi\cos\theta$. Similarly, $\hat{\mathbf{e}}_\rho\cdot\hat{\mathbf{j}} = y/\rho = \sin\phi\sin\theta$ and $\hat{\mathbf{e}}_\rho\cdot\hat{\mathbf{k}} = z/\rho = \cos\phi$. To obtain $\hat{\mathbf{e}}_\phi$ it is simplest to observe that it can be obtained from $\hat{\mathbf{e}}_\rho$ by adding $\pi/2$ to ϕ: $\hat{\mathbf{e}}_\phi = \sin(\phi + \pi/2)(\cos\theta\,\hat{\mathbf{i}} + \sin\theta\,\hat{\mathbf{j}}) + \cos(\phi + \pi/2)\,\hat{\mathbf{k}}$. And since the spherical coordinate base vector $\hat{\mathbf{e}}_\theta$ is the same

as the cylindrical coordinate base vector $\hat{\mathbf{e}}_\theta$ we have $\hat{\mathbf{e}}_\theta = -\sin\theta\,\hat{\mathbf{i}} + \cos\theta\,\hat{\mathbf{j}}$.
4. (a) $d\mathbf{A}/dt = (3\cos 3t + 27t^4)\,\hat{\mathbf{e}}_r + (3\sin 3t - 36t^3)\,\hat{\mathbf{e}}_\theta + 2t\,\hat{\mathbf{e}}_z$ (d) $d\mathbf{A}/dt = 2t\,\hat{\mathbf{e}}_\rho + 2t^2\,\hat{\mathbf{e}}_\phi + 3t^2\sin 2t\,\hat{\mathbf{e}}_\theta$
5. (a) $\mathbf{v} = \dot{x}\,\hat{\mathbf{e}} + \dot{y}\,\hat{\mathbf{j}} + \dot{z}\,\hat{\mathbf{k}} = 2t\,\hat{\mathbf{i}} = 2t(\cos\theta\,\hat{\mathbf{e}}_r - \sin\theta\,\hat{\mathbf{e}}_\theta) = 2t\left(\frac{x}{r}\,\hat{\mathbf{e}}_r - \frac{y}{r}\,\hat{\mathbf{e}}_\theta\right) = \frac{2t}{\sqrt{t^4+4}}(t^2\,\hat{\mathbf{e}}_r - 2\,\hat{\mathbf{e}}_\theta)$, $\mathbf{a} = \dot{\mathbf{v}} = 2\,\hat{\mathbf{i}} = \frac{2}{\sqrt{t^4+4}}(t^2\,\hat{\mathbf{e}}_r - 2\,\hat{\mathbf{e}}_\theta)$ **6.** (a) $\mathbf{v} = \dot{x}\,\hat{\mathbf{j}} + \dot{y}\,\hat{\mathbf{j}} + \dot{z}\,\hat{\mathbf{k}} = 2t\,\hat{\mathbf{i}} = 2t(\sin\phi\cos\theta\,\hat{\mathbf{e}}_\rho + \cos\phi\cos\theta\,\hat{\mathbf{e}}_\phi - \sin\theta\,\hat{\mathbf{e}}_\theta) = 2t\left(\frac{x}{\rho}\,\hat{\mathbf{e}}_\rho + \frac{y}{\rho}\,\hat{\mathbf{e}}_\phi + \frac{z}{\rho}\hat{\mathbf{e}}_\theta\right) = \frac{2}{\sqrt{t^4+4}}\left(t^2\,\hat{\mathbf{e}}_\rho + 2\,\hat{\mathbf{e}}_\phi\right)$, $\mathbf{a} = \dot{\mathbf{v}} = 2\,\hat{\mathbf{i}} = \frac{2}{\sqrt{t^4+4}}\left(t^2\,\hat{\mathbf{e}}_\rho + 2\,\hat{\mathbf{e}}_\phi\right)$ **9.** (c) $\rho = s = Vt$, $\theta = \Omega t$, $\phi = \pi/4$ so (30) and (31) give $\mathbf{v}(t) = V\,\hat{\mathbf{e}}_\rho + \frac{V}{\sqrt{2}}\,\hat{\mathbf{e}}_\theta$, $\mathbf{a}(t) = -\frac{1}{2}V\Omega^2 t(\,\hat{\mathbf{e}}_\rho + \hat{\mathbf{e}}_\phi) + \sqrt{2}\,V\Omega\,\hat{\mathbf{e}}_\theta$

Chapter 15

Section 15.2

1. (a) Gauss elimination gives the solution set as $z = \alpha$, $y = (3\alpha - 8)/5$, $x = (\alpha + 4)/5$ so we have a parametrization $x = (\tau + 4)/5$, $y = (3\tau - 8)/5$, $z = \tau$ $(-\infty < \tau < \infty)$ (f) $x^2 + y^2 = 4$ suggests letting $x = 2\cos\tau$, $y = 2\sin\tau$ $(0 \le \tau < 2\pi)$. Then $x + y + 2z = 5$ gives $z = (5 - x - y)/2 = \frac{5}{2} - \cos\tau - \sin\tau$. **2.** (a) $x = 5 - 3\tau$, $y = -1 + \tau$, $z = 2 + 4\tau$ $(0 \le \tau \le 1)$ **3.** (a) $R'(t)\cdot R'(t) = 1 + 4\tau^2$ so $s(\tau) = \int_0^\tau \sqrt{1+4t^2}\,dt = \frac{1}{2}\tau\sqrt{1+4\tau^2} + \frac{1}{4}\ln(2\tau + \sqrt{1+4\tau^2})$ **11.** (a) Det = 0 so the curve is a plane curve.

Section 15.3

1. (a) $\int_0^1\int_0^y y^2\,dxdy = \int_0^1\int_x^1 y^2\,dydx = 1/4$
(d) $\int_1^2\int_0^y \sin(x-y)\,dxdy = \int_0^1\int_1^2 \sin(x-y)\,dydx + \int_1^2\int_x^1 \sin(x-y)\,dydx = \sin 2 - \sin 1 - 1$
4. (a) $M = \sigma ab/2$, $x_c = 2a/3$, $y_c = b/3$ (b) $M = 4\sigma$, $x_c = 14/15$, $y_c = 11/5$ **5.** (a) $I_x = 3\sigma/4$, $I_y = 65\sigma/12$ (b) $I_x = 46\sigma/3$, $I_y = 10\sigma/3$ **10.** (a) -14 **11.** 1/3

Section 15.4

8. (a) $x = u$, $y = v$ and $x = u$, $y = u + v$ $(-\infty < u < \infty, -\infty < v < \infty)$ (d) $z = u$, $y = v$, $x = 2 - \frac{1}{3}u + \frac{2}{3}v$ $(-\infty < u < \infty, -\infty < v < \infty)$ **11.** (a) (25) gives $\hat{\mathbf{n}} = \pm(4\hat{\mathbf{i}} + 2\hat{\mathbf{j}} - \hat{\mathbf{k}})/\sqrt{21}$ (d) $\hat{\mathbf{n}} = \pm(\hat{\mathbf{i}} + 3\hat{\mathbf{j}} + \hat{\mathbf{k}})/\sqrt{11}$
(g) $\mathbf{n} = \mathbf{R}_u \times \mathbf{R}_v = \hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}}$, $\hat{\mathbf{n}} = \pm(\hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}})/\sqrt{18}$

Section 15.5

1. (a) With $E = 1$, $F = 0$, $G = 1$, (5) gives $dA = dudv$ so $A = \int_0^{2\pi}\int_0^{1+\sin v} dudv = 2\pi$. (c) With $E = 2$, $F = 0$, $G = u^2$, (5) gives $dA = \sqrt{2}\,u\,dudv$ so $A = \int_0^{2\pi}\int_0^h \sqrt{2}\,u\,dudv = \sqrt{2}\pi h^2$. **6.** (a) $\pi/4$ **10.** (a) With $z = 1 + y$, (18) gives $dA = \sqrt{1+0+1}\,dxdy = \sqrt{2}\,dxdy$. Hence, $\int_0^1\int_0^1 (1+x)\sqrt{2}\,dxdy = 3/\sqrt{2}$. (c) $2\pi h$
(e) With $x = 3\sin v\cos u$, $y = 3\sin v\sin u$, $z = 3\cos v$, we obtain $\int_0^\pi\int_{-\pi/2}^{\pi/2}(1 + 3\sin v\cos u)9\sin v\,dudv = 45\pi$.
(g) $\sqrt{2}\pi h^2$

Section 15.6

3. $\mathbf{F}(0,0,0) = \pi RG\sigma\hat{\mathbf{k}}$ **4.** (a) $z_c = 3h/4$ **8.** (a) $\mathbf{F} = 2\pi\sigma Gh\hat{\mathbf{k}}$ on $z = 0$, $\mathbf{F} = -2\pi\sigma Gh\hat{\mathbf{k}}$ on $z = h$

Chapter 16

Section 16.2

4. (a) $\mathbf{w}(2,3,-1) = 3\hat{\mathbf{i}} - 2\hat{\mathbf{j}} - \hat{\mathbf{k}}$ so we want $\frac{y}{3} = \frac{-x}{-2} = \frac{z}{-1}$. Setting $x = \tau$, say, gives the parametric equations of the desired curve as $x = \tau$, $y = 3\tau/2$, $z = -\tau/2$ $(-\infty < \tau < \infty)$. (d) $x = 2$, $y = \tau$, $z = -3/\tau$ $(0 < \tau < \infty)$. To understand our choice of the τ interval, note that $z\tau = -3$ gives a hyperbola in the τ, z plane, with one branch in the fourth quadrant and one in the second. The one in the fourth quadrant passes through the given point $z = -1$, so we choose that branch, for which $0 < \tau < \infty$.

Section 16.3

1. (a) div $\mathbf{v} = 0$ everywhere (d) div $\mathbf{v} = yz + xz - 3xy = 17$ at $(3, -1, 4)$

Section 16.4

1. (a) grad $u = 6\hat{\mathbf{i}}$ everywhere (e) grad $u = yz\hat{\mathbf{i}} + xz\hat{\mathbf{j}} + xy\hat{\mathbf{k}} = -4\hat{\mathbf{i}} - 9\hat{\mathbf{j}} + 36\hat{\mathbf{k}}$ **2.** (a) $du/ds = \nabla u \cdot \hat{\mathbf{v}} = (2x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} + 2z\hat{\mathbf{k}}) \cdot \hat{\mathbf{i}} = 2x = 4$ at (2,1,5) (d) $11/\sqrt{3}$ **4.** (a) $\boldsymbol{\nabla} V = (6xy - z)\hat{\mathbf{i}} + 3x^2\hat{\mathbf{j}} - x\hat{\mathbf{k}} = 37\hat{\mathbf{i}} + 12\hat{\mathbf{j}} - 2\hat{\mathbf{k}}$ is in the direction of maximum rate of *increase*, so the charge will move in the direction $-37\hat{\mathbf{i}} - 12\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$.
6. (a) $dT/dt = 200t + (20xt\hat{\mathbf{i}} - 10y\hat{\mathbf{j}}) \cdot (4x\hat{\mathbf{i}} + 4y\hat{\mathbf{j}}) = 2,040$ at $(2, -1, 3, 4)$

Section 16.5

1. (a) curl $\mathbf{v} = \mathbf{0}$ everywhere (d) curl $\mathbf{v} = x(z - y)\hat{\mathbf{i}} - y(z - x)\hat{\mathbf{j}} + z(y - x)\hat{\mathbf{k}} = -15\hat{\mathbf{i}} + 16\hat{\mathbf{j}} - \hat{\mathbf{k}}$ at $(3, 4, -1)$

Section 16.6

1. (a) $0, \mathbf{0}$ (d) $2y^3 + 6x^2y, \mathbf{0}$ **2.** (a) $0, \mathbf{0}$ (d) $0, 2\hat{\mathbf{j}}$ **7.** (d) $\boldsymbol{\nabla} \times \mathbf{H} = \mathbf{J}$ **11.** (a) $\nabla^2\mathbf{v} = \nabla^2(xz^2)\hat{\mathbf{i}} + \nabla^2(y \sin z)\hat{\mathbf{k}} = 2x\hat{\mathbf{i}} - y \sin z\hat{\mathbf{k}}$

Section 16.7

1. (a) $\boldsymbol{\nabla} u = \hat{\mathbf{e}}_r$, $\nabla^2 u = 1/r$, $\boldsymbol{\nabla} \cdot \mathbf{v} = 3$ [which makes sense because $\mathbf{v} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$ is the position vector to the point and is therefore expressible in Cartesian coordinates as $x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, with divergence $\partial(x)/\partial x + \partial(y)/\partial y + \partial(z)/\partial z = 3$], $\boldsymbol{\nabla} \times \mathbf{v} = \mathbf{0}$. NOTE: In using (16) and (17) we have $v_r = r$, $v_\theta = 0$, $v_z = z$.
(d) $\boldsymbol{\nabla} u = z^3\hat{\mathbf{e}}_r + 3z^2r\hat{\mathbf{e}}_z$, $\nabla^2 u = 6zr + z^3/r$. With $v_r = v_z = 0$ and $v_\theta = 1$, (16) and (17) give $\boldsymbol{\nabla} \cdot \mathbf{v} = 0$ and $\boldsymbol{\nabla} \times \mathbf{v} = (1/r)\hat{\mathbf{e}}_z$. **2.** (a) $1/r$ (d) $-(2 \sin\theta \cos\theta)/r$ **8.** (a) $3\hat{\mathbf{e}}_r$ **11.** (a) $\boldsymbol{\nabla} u = \hat{\mathbf{e}}_\rho$, $\nabla^2 u = 2/\rho$, $\boldsymbol{\nabla} \cdot \mathbf{v} = 3$, $\boldsymbol{\nabla} \times \mathbf{v} = \mathbf{0}$ (d) $\boldsymbol{\nabla} u = 2\rho \sin\theta\hat{\mathbf{e}}_\rho + \rho(\cos\theta/\sin\phi)\hat{\mathbf{e}}_\theta$, $\nabla^2 u = \left[6 - (1/\sin^2\phi\right] \sin\theta$, $\boldsymbol{\nabla} \cdot \mathbf{v} = (2 + \cot\phi)/\rho$, $\boldsymbol{\nabla} \times \mathbf{v} = (\cot\phi\,\hat{\mathbf{e}}_\rho - \hat{\mathbf{e}}_\phi + \hat{\mathbf{e}}_\theta)/\rho$

Section 16.8

1. (a) Each is 0 (d) Each is 1/3 (g) Each is 1/2 **4.** (a) Each is $2ab$ **5.** (a) Use $\mathbf{v} = (x^3yz/3)\hat{\mathbf{i}}$, for instance
12. (a) Each is 1920π (g) Each is $2\pi a^5$

Section 16.9

2. (a) Each is 1/3 (e) Each is $-1/6$ (i) Each is 1/12 **3.** (a) Each is -2/15 (e) Each is 0 **11.** (a) Each is -63
16. (a) Each is $2\pi a^2\omega$

Section 16.10

1. (a) All of 3-space (e) Everywhere except along the z axis ($x = y = 0$) **2.** (a) $\Phi = x - 2y - 8z$; everywhere
(f) $\Phi = 2ye^x - 3z$ in $x < 0$ (k) Irrotational nowhere **3.** (a) No **5.** (c) $\Phi = 2x^{5/2}e^{2y} + \frac{5}{3}y^3$, $I = -\frac{10}{3}$
6. (a) $\Phi = (x^2 + y^2 + z^2)/2$, $I = \pi^2(\pi^2 + 9)/2$ **11.** (a) $\boldsymbol{\nabla} \cdot \mathbf{v} = 0 + 0 + 0 = 0$. With $x_0 = y_0 = 0$, say, (10.3) gives $\mathbf{w} = cx\hat{\mathbf{j}} + (ay - bx)\hat{\mathbf{k}}$, to which we can add the gradient of an arbitrary function f. (d) $\mathbf{w} = -xyz\,\hat{\mathbf{j}}$, to which we can add the gradient of an arbitrary function f. **13.** (a) $\Phi = rz$, $I = 2$ (e) $\Phi = \rho^6/6$, $I = 62/3$

(h) $\Phi = \sin\phi + \cos\theta$, $I = 2(\sqrt{5} - 5)/5$

Chapter 17

Section 17.2

5. (a) $f_e(x) = 2$, $f_o(x) = -5x$ (e) $f_e(x) = x^2/(x^2-4)$, $f_o(x) = 2x/(4-x^2)$ **9.** $f_e(x) + f_o(x) = g_e(x) + g_o(x)$ and, changing x to $-x$, $f_e(-x) + f_o(-x) = g_e(-x) + g_o(-x)$ or $f_e(x) - f_o(x) = g_e(x) - g_o(x)$. Adding these two equations gives $2f_e(x) = 2g_e(x)$, and subtracting gives $2f_o(x) = 2g_o(x)$. Hence $f_e(x) = g_e(x)$ and $f_o(x) = g_o(x)$. **12.** (a) No (g) Yes, π (m) Yes, $2\pi/3$ (s) No (Sketch its graph) **13.** (a) 2π (e) 2π

Section 17.3

1. (a) Yes, in fact it is continuous (d) No, its limit does not exist as $x \to 0$ **4.** (a) $\text{FS}f = 2\sum_1^\infty \left[(-1)^{n+1}/n\right] \sin nx$ converges to $f(x)$ at all x except at $\pm\pi, \pm 3\pi, \pm 5\pi, \dots$, where $f(x)$ is π but the series converges to the average value 0 (d) $\text{FS}f = 50$ is identical to $f(x)$ for all x (g) $\text{FS}f = \frac{2}{\pi} - \frac{4}{\pi}\sum_1^\infty \left[1/(4n^2-1)\right] \cos 2nx$ converges to $f(x)$ for all x **5.** (a) $\text{FS}f = \pi^2/3 + \sum_1^\infty \left[(-1)^n/n^2\right] \cos nx$ **12.** (a) $l = 3$, $p(3) = 3$, $p'(3) = 1$, $p''(3) = \cdots = 0$, $b_n = 2(3)(-1)^{n+1}/(n\pi)$ so $\text{FS}f = \frac{6}{\pi}\sum_1^\infty \left[(-1)^{n+1}/n\right] \sin\frac{n\pi x}{3}$ (d) $\text{FS}f = -\frac{112}{15} + \dfrac{768}{\pi^4}\displaystyle\sum_1^\infty \left[(-1)^{n+1}/n^4\right] \cos\dfrac{n\pi x}{2}$ **16.** (a) $\text{FS}f = 25 + \frac{50}{\pi}\sum_{-\infty}^{\infty} \frac{1}{n} \sin\frac{n\pi}{2} e^{in\pi x/2}$ (d) $\text{FS}f = 6(e^{ix} - e^{-ix})/2i = 3ie^{-ix} - 3ie^{ix}$ (a 2-term series) **18.** (a) $x(t) = 50 + \frac{800}{\pi}\sum_{n=1,3,\dots}^{\infty} \dfrac{1}{n(4-n^2\pi^2)} \sin\dfrac{n\pi t}{2}$ (d) $x(t) = \frac{5}{2} - \dfrac{20}{\pi^2}\displaystyle\sum_{n=1,3,\dots}^{\infty} \dfrac{1}{n^2(1-n^2\pi^2)} \cos n\pi t$

Section 17.4

2. (a) HRC: $f(x) = \frac{25}{2} + \frac{50}{\pi}\sum_1^\infty \frac{1}{n}\sin\frac{n\pi}{2}\cos\frac{n\pi x}{2}$, HRS: $f(x) = \frac{50}{\pi}\sum_1^\infty \frac{1}{n}(1 - \cos\frac{n\pi}{2})\sin\frac{n\pi x}{2}$, QRC: $f(x) = \frac{100}{\pi}\sum_{1,3,\dots}^\infty \frac{1}{n}\sin\frac{n\pi}{4}\cos\frac{n\pi x}{4}$, QRS: $f(x) = \frac{100}{\pi}\sum_{1,3,\dots}^\infty \frac{1}{n}(1 - \cos\frac{n\pi}{4})\sin\frac{n\pi x}{4}$

Section 17.5

2. (a) $|e^{-nx}\sin nx| < e^{-2n}$ on $2 < x < 5$, and $\sum_1^\infty e^{-2n}$ is convergent [by the ratio test or by writing it as the geometric series $\sum_1^\infty (e^{-2})^n$, which converges because $|e^{-2}| < 1$]. Hence, the given series is uniformly convergent on $2 < x < 5$ by Theorem 17.5.1. **3.** (a) On $1 < x_0 < x < \infty$, $\left|\dfrac{1}{1+x^n}\right| \le \dfrac{1}{1+x_0^n} \sim \dfrac{1}{x_0^n} = (1/x_0)^n$, and $\sum_0^\infty (1/x_0)^n$ is a convergent geometric series if $x_0 > 1$. Hence, the series converges uniformly on $x_0 < x < \infty$ for any $x_0 > 1$, by Theorem 17.5.1. (d) On $|x| < x_0$ for any $x_0 < 1/5$ **4.** (a) $\frac{d}{dx}\sum_1^\infty \dfrac{\sin 2nx}{n^4} = \displaystyle\sum_1^\infty \dfrac{d}{dx}\left(\dfrac{\sin 2nx}{n^4}\right) = 2\sum_1^\infty \dfrac{\cos 2nx}{n^3}$ according to Theorem 17.5.2 because the latter series converges uniformly by Theorem 17.5.1, with $M_n = 1/n^3$. (Recall from the calculus that the ***p*-series** $\sum_1^\infty 1/n^p$ converges if $p > 1$; in this case $p = 3 > 1$.) **5.** (a) Expanding $f(t)$, let us call $x'' + x' + x = \frac{2}{\pi} - \frac{4}{\pi}\sum_1^\infty \dfrac{1}{4n^2-1}\cos 2nt$ equation (A). A particular solution of (A) due to the $2/\pi$ is $x_p = 2/\pi$, and a particular solution of $x'' + x' + x = \cos 2nt$ is $x_p(t) = [(1-4n^2)^2 \cos 2nt + 2n \sin 2nt]/[(1-4n^2)^2 + 4n^2]$ so, by superposition, a particular solution of (A) is $x_p(t) = \frac{2}{\pi} - \frac{4}{\pi}\sum_1^\infty \dfrac{1}{4n^2-1}\dfrac{(1-4n^2)\cos 2nt + 2n\sin 2nt}{(1-4n^2)^2 + 4n^2}$. The latter does satisfy (A), as can be verified by substitution, provided that we can justify the termwise differentiations of the series needed for x' and x''. Differentiating termwise gives $x_p'(t) = -\frac{4}{\pi}\sum_1^\infty \dfrac{1}{4n^2-1}\dfrac{-2n(1-4n^2)\sin 2nt + 4n^2\cos 2nt}{(1-4n^2)^2 + 4n^2} \equiv -\dfrac{4}{\pi}\displaystyle\sum_1^\infty a_n(t)$. Using

$|\sin 2nt| \leq 1$ and $|\cos 2nt| \leq 1$, $|a_n(t)| \leq \dfrac{2n + 8n^3 + 4n^2}{(4n^2 - 1)\left[(1 - 4n^2)^2 + 4n^2\right]} \sim \dfrac{1}{8n^3}$ as $n \to \infty$, so there will exist some constant C such that $|a_n(t)| \leq C/n^3$ for all $n = 1, 2, \ldots$. Since $\sum_1^\infty (C/n^3) = C \sum_1^\infty (1/n^3)$ converges, the termwise differentiation is justified by Theorem 17.5.1. Similarly for x''.

Section 17.6

2. (a) $\|\mathbf{E}\|^2 = \dfrac{2\pi^3}{3} - \pi\left[\dfrac{\pi^2}{2} + \dfrac{4}{\pi^2}\displaystyle\sum_1^k [(-1)^n - 1]^2/n^4\right]$ so $\|\mathbf{E}\| = 0.27, 0.27, 0.11, 0.11, 0.06, 0.06, 0.04, 0.04$ for $k = 1, 2, \ldots, 8$.

Section 17.7

1. (a) $\lambda_n = [(2n-1)\pi/2L]^2$, $\phi_n(x) = \sin[(2n-1)\pi x/2L]$ for $n = 1, 2, \ldots$, $100 = \frac{400}{\pi} \sum_1^\infty \frac{1}{2n-1} \sin \frac{2n-1}{2L}\pi x$ (c) $\lambda_n = (n\pi/L)^2$, $\phi_n(x) = \cos(n\pi x/L)$ for $n = 0, 1, \ldots$, $f(x) = \frac{1}{2} + \frac{2}{\pi}\sum_1^\infty \dfrac{(-1)^{n+1}}{2n-1} \cos[(2n-1)\pi x/L]$ **4.** (a) $\sigma(x) = x^4$, $(x^5 y')' + \lambda x^5 y = 0$, $p(x) = x^5$, $q(x) = 0$, $w(x) = x^5$ **9.** (a) $\lambda_n = n^2$, $\phi_n(x) = \cos nx$ for $n = 0, 1, 2, \ldots$, $f(x) = a_0 + \sum_1^\infty a_n \cos nx$, $a_0 = \langle f, 1\rangle/\langle 1, 1\rangle = \int_0^2 x^4\, dx/\pi = 32/(5\pi)$, $a_n = \langle f, \cos nx\rangle/\langle \cos nx, \cos nx\rangle = (2/\pi)\int_0^2 x^4 \cos nx\, dx = [32/(n^5\pi)][(2n^4 - 6n^2 + 3)\sin n \cos n + (4n^3 - 6n)\cos^2 n - 2n^3 + 3n]$ **17.** (a) $\langle L[u], v\rangle = \int_0^1 u'v\, dx = uv|_0^1 - \int_0^1 uv'\, dx$ so $L^*[v] = -v'$ (i.e., $L^* = -d/dx$), $v(1) = 0$, not Hermitian (c) $L^* = d^2/dx^2$, $v(1) = v'(1) = 0$, not Hermitian because the boundary conditions associated with L^* are different from those associated with L (e) $L^* = d^2/dx^2 + 3$, $v'(0) = v'(1) = 0$, Hermitian

Section 17.8

1. (a) $\lambda_0 = 0$, $\phi_0(x) = 1$; $\lambda_n = (n\pi/2)^2$, $\phi_n(x) = \cos(n\pi x/2)$ and $\sin(n\pi x/2)$ for $n = 1, 2, \ldots$, $H(x-2) = \frac{1}{2} - \frac{2}{\pi}\sum_{1,3,\ldots}^\infty \frac{1}{n}\sin\frac{n\pi x}{2}$ (c) $\lambda_0 = 0$, $\phi_0(x) = 1$; $\lambda_n = (2n\pi/\ln 2)^2$, $\phi_n(x) = \cos\left(2n\pi\frac{\ln x}{\ln 2}\right)$ for $n = 1, 2, \ldots$; expansions of the form $f(x) = a_0 + \sum_1^\infty a_n \cos\left(2n\pi\frac{\ln x}{\ln 2}\right)$, so if $f(x) = 6$ then $a_0 = 6$ and $a_n = 0$ for $n \geq 1$ by inspection. Perhaps that result is not obvious because the weight function is $w(x) = 1/x$, but $a_0 = \langle f(x), 1\rangle/\langle 1, 1\rangle = \int_1^2 (6)(1)x^{-1}\, dx / \int_1^2 (1)^2 x^{-1}\, dx = 6$ and the two integrals in the evaluation of a_n look difficult but are readily evaluated using the substitution $u = 2n\pi(\ln x)/\ln 2$. (e) See Section 4.4. $\lambda_n = n(n+1)$, $\phi_n(x) = P_n(x)$ for $n = 0, 2, 4, \ldots$; expansions of the form $f(x) = \sum_{0,2,4,\ldots}^\infty a_n P_n(x)$. If $f(x) = x$, then $a_0 = \langle x, 1\rangle/\langle 1, 1\rangle = 1/2$, $a_2 = \langle x, P_2(x)\rangle/\langle P_2(x), P_2(x)\rangle = 5/8$, $a_4 = \langle x, P_4(x)\rangle/\langle P_4(x), P_4(x)\rangle = -3/16$ **4.** From Section 4.4 we see that $\phi_n(x) = P_{2n}(x)$ for $n = 0, 1, 2, \ldots$, so $1 - x = \sum_0^\infty a_n P_{2n}(x)$ where $a_n = \langle 1 - x, P_{2n}(x)\rangle/\langle P_{2n}(x), P_{2n}(x)\rangle$. From (18) in Section 4.4, $\int_0^1 P_{2n}^2(x)\, dx = 1/(2n+1)$ so $\langle P_{2n}(x), P_{2n}(x)\rangle = \int_0^1 P_{2n}^2(x)\, dx = 1/(4n+1)$ and $a_n = (4n+1)\int_0^1 (1-x)P_{2n}(x)\, dx$: $a_0 = 1/2$, $a_1 = -5/8$, $a_2 = 9/48, \ldots$

Section 17.9

2. (a) $f(x) = (100/\pi)\int_0^\infty [\sin 2\omega \cos\omega x + (1 - \cos 2\omega)\sin\omega x]\, d\omega/\omega$ for all x except $x = 0$ and $x = 2$: $f(0) = f(2) = 100$ but the Fourier integral of f converges to the average value 50. As a check, application of the *Maple* int command gives these values for the Fourier integral, integrating from 0 to 5000: $200Si(5000)/\pi$ at $x = 1$ and $-100[Si(10,000) - Si(20,000)]/\pi$ at $x = 4$. Following the int command with evalf(") evaluates these quantities as 99.96 and −0.0043, respectively. (d) $f(x) = (1/\pi)\int_0^\infty [(\cos 5\omega + 5\omega \sin 5\omega - 1)\cos\omega x + (5\omega\cos 5\omega - \sin 5\omega)\sin\omega x]\, d\omega/\omega^2$ for all x except $x = -5$: $f(-5) = 0$ but the Fourier integral of f converges to 2.5.

Section 17.10

4. (a) $f(x) = a/[\pi(a^2 + x^2)]$ (b) $f(x) = \sin ax/(\pi x)$ **6.** (a) $F\left\{4x^2 e^{-3|x|}\right\} = 4i^2 \frac{d^2}{d\omega^2} F\left\{e^{-3|x|}\right\}$ (entry 17) $= -4\frac{d^2}{d\omega^2}\left(\frac{6}{\omega^2+9}\right)$ (entry 4) $= 48\left[\frac{1}{(\omega^2+9)^2} - \frac{4\omega^2}{(\omega^2+9)^3}\right]$ (c) With $c = 3$, $a = 1$, $f(x) = 1/(x^2 + 2)$ in entry 13, and $\hat{f} = (\pi/\sqrt{2})\exp(-\sqrt{2}|\omega|)$, we have $F\left\{\frac{\cos 3x}{x^2+2}\right\} = \frac{1}{2}[\hat{f}(\omega-3)+\hat{f}(\omega+3)] = \frac{\pi}{2\sqrt{2}}(e^{-\sqrt{2}|\omega-3|}+e^{-\sqrt{2}|\omega+3|})$ (e) $F\left\{\frac{3}{2x^2+1} - 5e^{-|x|}\right\} = \frac{3}{2}F\left\{\frac{1}{x^2+1/2}\right\} - 5F\left\{e^{-|x|}\right\}$ (entry 18) $= \frac{3}{2}\pi\sqrt{2}e^{-|\omega|/\sqrt{2}} - 5\frac{2}{\omega^2+1}$ (entries 1, 4) $= \frac{3\pi}{\sqrt{2}}e^{-|\omega|/\sqrt{2}} - \frac{10}{\omega^2+1}$ (g) $F^{-1}\left\{\frac{4\sin\omega}{\omega} - \frac{1}{\sqrt{|\omega|}}\right\} = 2F^{-1}\left\{\frac{2\sin\omega}{\omega}\right\} - \frac{1}{\sqrt{2\pi}}F^{-1}\left\{\sqrt{\frac{2\pi}{|\omega|}}\right\}$ (entry 18) $= 2[H(x+1) - H(x-1)] - \frac{1}{\sqrt{2\pi|x|}}$ (entries 9, 7) (i) $F^{-1}\left\{\frac{9}{2\omega+i}\right\} = -\frac{9i}{2}F^{-1}\left\{\frac{1}{1/2-i\omega}\right\}$ (entry 18) $= -\frac{9i}{2}H(-x)e^{x/2}$ (entry 3) (k) $F^{-1}\left\{e^{-|\omega|}\cos\omega\right\} = \frac{1}{2\pi}F^{-1}\left\{2(\pi e^{-|\omega|})\cos\omega\right\}$ (entry 18) $= \frac{1}{2\pi}\left[\frac{1}{(x+1)^2+1} + \frac{1}{(x-1)^2+1}\right]$ (entries 15, 1) (m) $F^{-1}\left\{\frac{1}{\omega^2+i\omega+2}\right\} = F^{-1}\left\{\frac{1}{3i}\left(\frac{1}{\omega-i} - \frac{1}{\omega+2i}\right)\right\} = F^{-1}\left\{\frac{1}{3}\left(\frac{1}{1+i\omega} - \frac{1}{-2+i\omega}\right)\right\} = \frac{1}{3}F^{-1}\left\{\frac{1}{1+i\omega} + \frac{1}{2-i\omega}\right\}$ (entry 18) $= \frac{1}{3}[H(x)e^{-x} + H(-x)e^{2x}]$ (entries 2, 3)

Section 17.11

9. (a) Because $u(0)$ is given, use the sine transform. Obtain $\hat{u}_S = -50\frac{\omega}{(\omega^2+9)^2}$ and consider entry 2S with $a = 3$. Since $\text{Im}(3+i\omega)^2 = 6\omega$, write $\hat{u}_S = -\frac{50}{6}\frac{6\omega}{(\omega^2+9)^2} = -\frac{25}{3}\frac{\text{Im}(3+i\omega)^2}{(\omega^2+3^2)^2}$ so $u(x) = -\frac{25}{3}xe^{-3x}$.

Chapter 18

Section 18.2

2. (a) $L = \nabla^2+k^2$ is linear because $L[\alpha u+\beta v] = (\nabla^2+k^2)(\alpha u+\beta v) = \alpha(\nabla^2+k^2)u+\beta(\nabla^2+k^2)u = \alpha L[u]+\beta L[v]$ (e) The operator defined by $L[u] = u_{xx} + u_{yy} - e^u$ is nonlinear because $L[\alpha u + \beta v] - \alpha L[u] - \beta L[v] = (\alpha u + \beta v)_{xx} + (\alpha u + \beta v)_{yy} - e^{(\alpha u+\beta v)} - \alpha(u_{xx} + u_{yy} - e^u) - \beta(v_{xx} + v_{yy} - e^v) = \alpha e^u + \beta e^v - e^{\alpha u}e^{\beta v} \neq 0$ in general. For example, if $\alpha = 2$, $\beta = 0$, $u(x,y) = x$, $v(x,y) = y$, then the latter is $2e^x - e^{2x}$, which is not identically zero (on any x, y domain of interest). **3.** (a) $A = 1$, $B = 1/2$, $C = 0$ so $B^2 - AC = 1/4 > 0$, so hyperbolic (everywhere in x, y plane). Note that D, E, F, and f are irrelevant insofar as this classification. (d) $B^2 - AC = 0 + x(\sin^2 y + 1)$, so hyperbolic in the half plane $x > 0$, elliptic in the half plane $x < 0$, parabolic on the line $x = 0$.

Section 18.3

4. (a) $X''T = XT' + 3XT$, $\frac{X''}{X} = \frac{T'}{T} + 3 = -\kappa^2$, $X'' + \kappa^2 X = 0$ and $T' + (\kappa^2 + 3)T = 0$ so $X = A\cos\kappa x + B\sin\kappa x$ and $T = C\exp(-\kappa^2 - 3)t$ for $\kappa \neq 0$, and $X = D + Ex$ and $T = Fe^{-3t}$ for $\kappa = 0$. Thus, $u(x,t) = (D + Ex)Fe^{-3t} + (A\cos\kappa x + B\sin\kappa x)C\exp(-\kappa^2 - 3)t = (G + Hx)e^{-3t} + (I\cos\kappa x + J\sin\kappa x)\exp(-\kappa^2 - 3)t$, say (c) $\frac{X''}{X} + 2\frac{X'T'}{XT} = \frac{T'}{T}$, which cannot be separated because of the mixed term $X'T'/XT$ **6.** (a) $u(x,t) = 20 + 3x + \sum_{1,3,\dots}^{\infty} A_n \sin\frac{nx}{2}\, e^{-(n\alpha/2)^2 t}$ is the solution, where (with the help of a quarter-range sine expansion) $A_n = (2/\pi)\int_0^\pi(-20 - 3x)\sin\frac{nx}{2}\,dx = \frac{2}{\pi}\left[\frac{40+6\pi}{n}\cos\frac{n\pi}{2} - \frac{12}{n^2}\sin\frac{n\pi}{2}\right]$. $u_s(x) = 20+3x$. (d) $u(x,t) = \sum_{1,3,\dots}^{\infty} A_n \sin\frac{n\pi x}{4}\, e^{-(n\pi\alpha/4)^2 t}$, where the A_n's are found from the initial condition $u(x,0) = 5\sin(\pi x/4) - 12\sin(5\pi x/4) = \sum_{1,3,\dots}^{\infty} A_n \sin\frac{n\pi x}{4}$ (on $0 < x < 2$). The quarter-range sine formula gives $A_n = \int_0^2 \left(5\sin\frac{\pi x}{4} - 12\sin\frac{5\pi x}{4}\right)\sin\frac{n\pi x}{4}\,dx = 10\left\{\frac{\sin[(n-1)\pi/2]}{(n-1)\pi} - \frac{\sin[(n+1)\pi/2]}{(n+1)\pi}\right\} - 24\left\{\frac{\sin[(n-5)\pi/2]}{(n-5)\pi} - \frac{\sin[(n+5)\pi/2]}{(n+5)\pi}\right\}$ as found using the *Maple* int command or tables. Observe that each sine term is zero because n $(= 1, 3, \dots)$ is odd, so A_n appears to be zero for each $n = 1, 3, \dots$. However, for $n = 1$ the $n - 1$ denominator vanishes and for $n = 5$ the $n - 5$ denominator vanishes so the latter expression is indeterminate for $n = 1$ and $n = 5$ (of the form 0/0). For those caes l'Hôpital's rule gives $A_1 = 10\left\{\frac{1}{2}\right\} = 5$ and $A_5 = -24\left\{\frac{1}{2}\right\} = -12$, with $A_n = 0$ for $n = 3, 7, 9, 11, 13, \dots$. Thus, the series solution reduces to the two-term result $u(x,t) = 5\sin\frac{\pi x}{4}\, e^{-(\pi\alpha/4)^2 t} - 12\sin\frac{5\pi x}{4} e^{-(5\pi\alpha/4)^2 t}$. Observe that this result is obtained more readily, from the initial condition, merely by matching terms: $5 = A_1$, $-12 = A_5$, and $0 = A_n$ for all other n's $(3, 7, 9, 11, \dots)$. $u_s(x) = 0$ (g) $u(x,t) = A + \sum_1^{\infty} B_n \cos nx\, e^{-(n\alpha)^2 t}$ and

$u(x,0) = 300 = A + \sum_1^\infty B_n \cos nx$ gives (half-range cosine formulas) $A = (1/\pi)\int_0^\pi 300\,dx = 300$, $B_n = (2/\pi)\int_0^\pi 300 \cos nx\,dx = (600 \sin n\pi)/(n\pi) = 0$ for each n. Thus, the solution is simply $u(x,t) = 300$. With a little experience, one could anticipate this result by inspection because the ends of the rod are insulated ($u_x = 0$ at each end) and the initial temperature distribution is uniform, so why should the temperature change? Or, mathematically, observe that $u(x,y) = 300$ does satisfy the PDE, the boundary conditions, and the initial condition.
7. Consider, for example, $t = 1000$. Running the *Maple* sum command for the sum on the right-hand side of (45), with $n = 1..20$, gives -43.28600284. Running it again, with $n = 1..30$, gives the same result so it is reasonable to assume that the result is correct to that many decimal places. However, understand that this reasoning is only heuristic. (In fact, it is interesting to run it for $n = 1..2$, $n = 1..4$, $n = 1..6$, and so on. We find that the sum settles down to 10 significant figures at $n = 1..8$; i.e., the series converges rapidly for this choice of x and t.) Finally, the command evalf(100 + ") gives the value 56.71399716 for $u(0,1000)$, which appears to agree with the value obtained (crudely) from Fig. 6.
10. (a) $u_s(x) = u_1 + Q_2 x$ (d) $u_s(x) = (u_2 - u_1 \cosh \beta L)\frac{\sinh \beta x}{\sinh \beta L} + u_1 \cosh \beta x$, where $\beta = \sqrt{H}/\alpha$ (h) $u_s(x) = A + Be^{Vx/\alpha^2}$, $A = (u_2 - u_1 e^{VL/\alpha^2})/(1 - e^{VL/\alpha^2})$, $B = (u_1 - u_2)/(1 - e^{VL/\alpha^2})$ **14.** (a) $a_0(t) = A_0$, $a_n(t) = A_n e^{-(2n\pi\alpha/L)^2 t}$, $b_n(t) = B_n e^{-(2n\pi\alpha/L)^2 t}$, where $u(x,0) = f(x) = A_0 + \sum_1^\infty \left(A_n \cos \frac{2n\pi x}{L} + B_n \sin \frac{2n\pi x}{L}\right)$ gives $A_0 = \frac{1}{L}\int_0^L f(x)\,dx$, $A_n = \frac{1}{L}\int_0^L f(x) \cos \frac{2n\pi x}{L}\,dx$, $B_n = \frac{1}{L}\int_0^L f(x) \sin \frac{2n\pi x}{L}\,dx$ **16.** (a) $u(x,t) = u_s(x) + \sum_1^\infty A_n \sin \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t}$, where $u_s(x) = Fx(L-x)/(2\alpha^2)$ and $A_n = -\frac{2}{L}\int_0^L u_s(x) \sin \frac{n\pi x}{L}\,dx$
17. (b) $u(x,t) = \frac{4}{\pi}\sum_{1,3,\dots}^\infty \frac{1}{n}\frac{1}{(n\pi\alpha/L)^2 - 1} \sin \frac{n\pi x}{L}\left(e^{-t} - e^{-(n\pi\alpha/L)^2 t}\right)$
30. (a) $u(x,t) = 50 + e^x \sum_{1,3,\dots}^\infty A_n \sin \frac{n\pi x}{L} e^{-[(n\pi/L)^2 + 1]t}$. Then $u(x,0) = 0 = 50 + \sum_{1,3,\dots}^\infty A_n e^x \sin \frac{n\pi x}{L}$. The eigenfunctions of the Sturm–Liouville problem $X'' - 2X' + \kappa^2 X = 0$ on $0 < x < L$ with $X(0) = 0$ and $X(L) = 0$ are $e^x \sin \frac{n\pi x}{L}$, so we satisfy $-50 = \sum_{1,3,\dots}^\infty A_n e^x \sin \frac{n\pi x}{L}$ by setting $A_n = \langle -50, e^x \sin \frac{n\pi x}{L}\rangle / \langle e^x \sin \frac{n\pi x}{L}, e^x \sin \frac{n\pi x}{L}\rangle$. To obtain the weight function for the inner product write $\sigma X'' - 2\sigma X' + \kappa^2 \sigma X = 0$, require that $-2\sigma = \sigma'$ and obtain $\sigma(x) = e^{-2x}$, which is also the weight function. Thus, $A_n = \int_0^L (-50) e^x \sin \frac{n\pi x}{L} e^{-2x}\,dx / \int_0^L \left(e^x \sin \frac{n\pi x}{L}\right)^2 e^{-2x}\,dx = -\frac{100}{L}\int_0^L e^{-x} \sin \frac{n\pi x}{L}\,dx$. NOTE We wrote $X(0) = 0$ and $X(L) = 0$ rather than $X(0) = 50$ and $X(L) = 50$, because the $u_s(x) = 50$ term in the solution satisfies the nonhomogeneous boundary conditions, so that the transient part of the solution satisfies a homogeneous, or "homogenized," version of those boundary conditions, which point was discussed in Example 3. Remember, a Sturm–Liouville problem is an eigenvalue problem so its boundary conditions *must* be homogeneous.

Section 18.4

8. (b) We can use either transform but the Laplace transform is a bit simpler because the transformed equation $\alpha^2 \bar{w}_{xx} - s\bar{w} = -\bar{F}(s)$ has only a constant forcing function [i.e., $\bar{F}(s)$ does not vary with the active variable x] whereas if we use the Fourier transform then the transformed equation $-\alpha^2\omega^2\hat{w} - \hat{w}_t = -F(t)$ has a nonconstant forcing function. Then $\bar{w}(x,s) = Ae^{\sqrt{s}x/\alpha} + Be^{-\sqrt{s}x/\alpha} + \bar{F}(s)/s$. Reasoning as in (29)–(31), we require that $\bar{w}(x,s) \to 0$ as $x \to \pm\infty$ so $A = B = 0$ and $w(x,t) = L^{-1}\{\bar{F}(s)/s\} = \int_0^t F(\tau)\,d\tau$ (which, as we could have anticipated from the beginning, varies with t but not with x). **15.** At $x = 0.5$ Hastings' formula gives erf $(0.5) \approx 0.5204876$, whereas tables give erf $(0.5) = 0.5204999$. The difference, 0.0000123, is indeed within the claimed accuracy.

Section 18.5

8. (a) Yes (d) No, due to the u^2 (g) Yes (j) Yes (m) No; $L = x^3\partial^2/\partial x^2 + 3x^2\partial/\partial x - \partial^2/\partial t^2 + 1$ and x^3 is not even, nor is $3x^2$ odd

Section 18.6

5. Here are a few: $U_{11} = 0.2$, $U_{21} = 0.2$, and $U_{33} = 0.44896$ **7.** Here are a few: $U_{11} = 76$, $U_{21} = 92$, and $U_{33} = 38.0096$, where we have used the average values $U_{00} = 50$ and $U_{40} = 50$ at the corners.
10. In the left half $r = r_L = 0.225$; in the right half $r = r_R = 0.025$. Thus, in the left half ($j = 1, 2$) use

$U_{j,k+1} = 0.225U_{j-1,k} + 0.55U_{jk} + 0.225U_{j+1,k}$ and in the right half ($j = 4, 5$) use $U_{j,k+1} = 0.025U_{j-1,k} + 0.95U_{jk} + 0.025U_{j+1,k}$. For the interface at $x = 6$ use $K_L(U_{3k} - U_{2k})/\Delta x = K_R(U_{4k} - U_{3k})/\Delta x$ or $U_{3k} = (K_L U_{2k} + K_R U_{4k})/(K_L + K_R) = 0.893U_{2k} + 0.107U_4$. Thus, $U_{11} = 11.25$, $U_{21} = U_{31} = U_{41} = U_{51} = 0$; $U_{12} = 28.69$, $U_{22} = 2.53$, $U_{32} = 0.893(2.53) + 0.107(0) = 2.26$, $U_{42} = U_{52} = 0$; $U_{13} = 38.85$, $U_{23} = 8.36$, then $U_{43} = 0.06$, then $U_{33} = 0.893(8.36) + 0.107(0.06) = 7.47$, $U_{53} = 0$; $U_{14} = 45.75$, $U_{24} = 15.02$, then $U_{44} = 0.24$, then $U_{34} = 0.893(15.02) + 0.107(0.24) = 13.44$, $U_{54} = 0.00$ **11.** (a) With exact values in parentheses, $U_{11} = U_{31} = 57.46\,(58.04)$, $U_{21} = 81.25\,(82.09)$, $U_{12} = U_{32} = 46.69\,(47.65)$, $U_{22} = 66.02\,(67.38)$, $U_{31} = U_{33} = 37.93\,(39.11)$, $U_{23} = 53.65\,(55.31)$

Chapter 19

Section 19.2

1. (a) $y(5,1) = -3/5$ (d) $y(5,4) = -3/5$ (g) $y(5,20) = 1$ **2.** (a) $y(x,t) = \frac{50L}{\pi c} \sin\frac{\pi x}{L} \sin\frac{\pi ct}{L}$. NOTE: The initial condition $y_t(x,0) = 50\sin\frac{\pi x}{L} = \sum_1^\infty \frac{n\pi c}{L} S_n \sin\frac{n\pi x}{L}$ gives, by the half-range sine formulas, $\frac{n\pi c}{L} S_n = \frac{2}{L}\int_0^L 50\sin\frac{\pi x}{L}\sin\frac{n\pi x}{L}\,dx$, which equals 50 for $n = 1$ and 0 for $n \geq 2$. However, it is much simpler to satisfy the initial condition (stated above) merely by matching coefficients of the $\sin(n\pi x/L)$ terms on the left- and right-hand sides.
5. (b) $y(x,t) = \sum_{1,3,\ldots}^\infty \sin\frac{n\pi x}{2L}\left(A_n\cos\frac{n\pi ct}{2L} + B_n\sin\frac{n\pi ct}{2L}\right)$. Then, $y(x,0) = f(x)$ gives
$A_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{2L}\,dx$ and $y_t(x,0) = 0$ gives $B_n = 0$, both by quarter-range sine series. **6.** $y(x,t) = e^{-at/2}\sum_1^\infty \sin\frac{n\pi x}{L}(A_n\cos\omega_n t + B_n\sin\omega_n t)$, where $\omega_n = \sqrt{(n\pi c/L)^2 - (a/2)^2}$. Then, $y(x,0) = f(x)$ gives $A_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx$ and $y_t(x,0)$ gives the B_n's in terms of the A_n's as $B_n = \frac{a}{2\omega_n}A_n$. The ay_y term in (6.1) causes these changes in the solution, all of which make sense physically: first, the $e^{-at/2}$ factor in the solution causes the motion to damp out as $t \to \infty$; second, $a > 0$ causes a reduction in the frequencies, from $\omega_n = n\pi c/L$ to $\omega_n = \sqrt{(n\pi c/L)^2 - (a/2)^2}$ (i.e., the damped system is more "sluggish"); third, $a > 0$ causes a phase shift, for if $a = 0$ then $B_n = 0$, and as a increases B_n increases. **8.** $y(x,t) = \frac{g}{2c^2}x(x-L) + \sum_1^\infty A_n\sin\frac{n\pi x}{L}\cos\frac{n\pi ct}{L}$, and $y(x,0) = f(x)$ gives $A_n = \frac{2}{L}\int_0^L[f(x) - \frac{g}{2c^2}x(x-L)]\sin\frac{n\pi x}{L}\,dx$. Note that even if $y(x,0) = f(x) = 0$ the solution is nonzero.
12. (b) $u(x,t) = \frac{8s_oL}{\pi^2E}\sum_{1,3,\ldots}^\infty \frac{1}{n^2}\sin\frac{n\pi}{2}\sin\frac{n\pi x}{2L}\cos\frac{n\pi ct}{2L}$ (c) $s(0,t) = E\epsilon(0,t) = Eu_x(0,t)$
$= \frac{4s_o}{\pi}\sum_{1,3,\ldots}^\infty \frac{1}{n}\sin\frac{n\pi}{2}\cos\frac{n\pi ct}{2L}$

Section 19.3

1. (a) $w(x,y,t)$ is given by (16a), where the H_{mn}'s are found from (17): $8\sin 2x\sin 2y = \cdots + H_{24}\sin\frac{2\pi x}{\pi}\sin\frac{4\pi y}{2\pi} + \cdots$ so, matching coefficients [which is much simpler than using (22)], we see that all the H_{mn}'s are zero except for H_{24}, which is 8. Thus, $w(x,y,t) = 8\sin 2x\sin 2y\cos 2\sqrt{2}t$. Nodal lines: $x = \pi/2$, $y = \pi/2,\ \pi,\ 3\pi/2$
(d) Matching coefficients in $\sin 3x\sin y - \sin x\sin 3y = \cdots + H_{32}\sin\frac{3\pi x}{\pi}\sin\frac{2\pi y}{2\pi} + H_{16}\sin\frac{1\pi x}{\pi}\sin\frac{6\pi y}{2\pi} + \cdots$ gives $H_{32} = 1$, $H_{16} = -1$ and all other H_{mn}'s are zero. Thus, $w(x,y,t) = (\sin 3x\sin y - \sin x\sin 3y)\cos\sqrt{10}\,t$. Nodal curve(s) are defined by the relation $\sin 3x\sin y - \sin x\sin 3y = 0$. **2.** $w(x,y,t) = 20\sin 3\pi x\sin 4\pi y\cos 5\pi t - 8\sin 5\pi x\sin 12\pi y\cos 13\pi t$. If the period is T, then for any integers p and q we have $5\pi T = 2p\pi$ and $13\pi T = 2q\pi$, so $T = 2p/5$ and $T = 2q/13$. To find the *fundamental* period we seek the smallest p and q such that $2p/5 = 2q/13$ or $p/q = 5/13$. They are $p = 5$ and $q = 13$, so the fundamental period is $T = 2$. **3.** You better believe it.

Section 19.4

2. (a) The change of variables reduces the PDE to the canonical form $u_{\xi\eta} = 0$, with general solution $u(\xi,\eta) = F(\xi) + G(\eta)$ or $u(x,y) = F(x-y) + G(3x-y)$, where F and G are arbitrary twice-differentiable functions. Comparing the PDE with (10) in Section 18.2 gives $A = 1$, $B = 2$, $C = 3$, so $B^2 - AC = 4 - 3 > 0$ and the PDE is hyperbolic. **3.** (a) The PDE becomes $(a^2+8ab+12b^2)u_{\xi\xi}+(2ac+8ad+8bc+24bd)u_{\xi\eta}+(c^2+8cd+12d^2)u_{\eta\eta}=0$. Setting $a^2+8ab+12b^2=0$ gives $a/b=-2$ or -6, and setting $c^2+8cd+12d^2=0$ gives $c/d=-2$ or -6 (the same as a/b by coincidence). Set $a=-2$, $b=1$, $c=-6$, $d=1$, say. Then $u_{\xi\eta}=0$ and $u(\xi,\eta)=F(\xi)+G(\eta)$ so $u(x,y)=F(-2x+y)+G(-6x+y)$. Comparing the PDE with (10) in Section 18.2 gives $B^2-AC=16-12>0$ so the PDE is hyperbolic. **7.** (a) It is easy to construct particular solutions – such as $-Kx^2/(2c^2)$ and $Kt^2/2$. Choosing the former, we can write the general solution as $y(x,t)=-Kx^2/(2c^2)+F(x-ct)+G(x+ct)$. (b) Imposing the initial conditions, we can solve for F and G but it is simpler to set $y(x,t)=-Kx^2/(2c^2)+v(x,t)$. Then v satisfies the problem $c^2v_{xx}=v_{tt}$; $v(x,0)=p(x)+Kx^2/(2c^2)$, $v_t(x,0)=q(x)$ and we can use the d'Alembert solution (17) to solve for $v(x,t)$. The result is $y(x,t)=-Kx^2/(2c^2)+\frac{1}{2}[P(x-ct)+P(x+ct)]+\frac{1}{2c}\int_{x-ct}^{x+ct}q(x')\,dx'$, where $P(x)=p(x)+Kx^2/(2c^2)$. **8.** $y(x,t)=H(t-\frac{x}{c})h(t-\frac{x}{c})$. NOTE: $y(x,t)=F(x-ct)+G(x+ct)$ gives $y(x,0)=0=F(x)+G(x)$ and $y_t(x,0)=0=-cF'(x)+cG'(x)$. Solving these gives $F(x)=\text{constant}\equiv A$, say, and $G(x)=-A$, so $y(x,t)=A-A=0$. The key step is to realize that the two boundary condition equations hold only for $x>0$, so $F(x)=0$ (we can let $A=0$ without loss) and $G(x)=0$ hold only for $x>0$. Thus, $F(x-ct)=0$ for $x>ct$ (the wedge $0<\theta<\pi/4$ in the first quadrant of the x,t plane) and $G(x+ct)=0$ for $x>-ct$ (the wedge $0<\theta<3\pi/4$), so $y(x,t)=0$ in $0<\theta<\pi/4$ and $y(x,y)=F(x-ct)$ in $\pi/4<\theta<\pi/2$. Then the boundary condition $y(0,t)=h(t)=F(-ct)$ gives $F(arg)=h\left(\frac{arg}{-c}\right)$, where "arg" denotes the argument of F, so $y(x,t)=F(x-ct)=h\left(\frac{x-ct}{-c}\right)$ holds in the wedge $\pi/4<\theta<\pi/2$.

16. (b) $(Y_{i0}-Y_{i,-1})/\Delta t=g(i\Delta x)$ gives $Y_{i,-1}=Y_{i0}-g(i\Delta x)\Delta t$. Hence, $Y_{1,-1}=0-\left(\sin\frac{\pi}{4}\right)(0.02)=-0.0141$, $Y_{2,-1}=0-\left(\sin\frac{\pi}{2}\right)(0.02)=-0.02$, $Y_{3,-1}=0-\left(\sin\frac{3\pi}{4}\right)(0.02)=-0.0141$. Then $Y_{11}=0.64Y_{00}+0.72Y_{10}+0.64Y_{20}-Y_{1,-1}=0.0141$, $Y_{21}=0.64Y_{10}+0.72Y_{20}+0.64Y_{30}-Y_{2,-1}=0.02$, $Y_{31}=0.64Y_{20}+0.72Y_{30}+0.64Y_{40}-Y_{3,-1}=0.0141$. Similarly, $Y_{12}=0.64Y_{01}+0.72Y_{11}+0.64Y_{21}-Y_{10}=0.0230$, $Y_{22}=0.64Y_{11}+0.72Y_{21}+0.64Y_{31}-Y_{20}=0.0324$, and $Y_{32}=0.64Y_{21}+0.72Y_{31}+0.64Y_{41}-Y_{30}=0.023$

Chapter 20

Section 20.2

1. (a) $u(x,y)=50\sin(\pi x/3)\left[\cosh(\pi y/3)-\coth(2\pi/3)\sinh(\pi y/3)\right]$. NOTE: Our results can sometimes be simplified using identities such as $\sinh(A\pm B)=\sinh A\cosh B\pm\cosh A\sinh B$ and $\cosh(A\pm B)=\cosh A\cosh B\pm\sinh A\sinh B$. Using the former, we can express our solution, alternatively, as $u(x,y)=50\sin\frac{\pi x}{3}\frac{\sinh[(2-y)\pi/3]}{\sinh(2\pi/3)}$.
(d) $u(x,y)=\sum_1^\infty A_n\sin\frac{n\pi x}{3}\left(\sinh\frac{n\pi y}{3}-\tanh\frac{2n\pi}{3}\cosh\frac{n\pi y}{3}\right)$, where (by half-range sine formulas) $u(x,0)=\sum_1^\infty -A_n\tanh\frac{2n\pi}{3}\sin\frac{n\pi x}{3}$ gives $-A_n\tanh\frac{2n\pi}{3}=\frac{2}{3}\int_0^3 u(x,0)\sin\frac{n\pi x}{3}\,dx$ or
$A_n=-\frac{2}{3}\coth\frac{2n\pi}{3}\int_2^3 50\sin\frac{n\pi x}{3}\,dx=\frac{100}{n\pi}\coth\frac{2n\pi}{3}[(-1)^n-\cos\frac{2n\pi}{3}]$
(h) $u(x,y)=\sum_{1,3,\dots}^\infty A_n\left(\sinh\frac{n\pi x}{4}-\tanh\frac{3n\pi}{4}\cosh\frac{n\pi x}{4}\right)\sin\frac{n\pi y}{4}$, where (by quarter-range sine formulas) $A_n=-\frac{4}{n\pi}\cos\frac{n\pi}{4}\coth\frac{3n\pi}{4}$ (k) $u(x,y)=\frac{5}{3\pi}(\sinh 3\pi x-\tanh 9\pi\cosh 3\pi x)\sin 3\pi y$ **2.** (a) $u(2.5,1)=6.07$ (4 terms), $u(2.5,0.5)=17.43$ (10 terms), $u(2.5,0.2)=33.87$ (27 terms), $u(2.5,0.1)=41.60$ (51 terms). Why are more terms needed as $y\to 0$? When $y=0$ then $u(x,0)=50H(x-2)$ is discontinuous so its Fourier series will converge very slowly indeed. As y increases from 0 the solution $u(x,y)$ becomes a more gradually varying function of x, so the series becomes more rapidly convergent. We urge you to use the series to evaluate $u(2.5,0)$, i.e., right *on* the boundary. How many terms are needed to achieve two-decimal-place accuracy? **4.** $u(x,y)=u_1+(u_2-u_1)\frac{y}{b}+\sum_1^\infty(A_n\sinh\frac{n\pi x}{b}+B_n\cosh\frac{n\pi x}{b})\sin\frac{n\pi y}{b}$, where B_n is given by $B_n=\frac{2}{b}\int_0^b[p(y)-u_1-(u_2-u_1)\frac{y}{b}]\sin\frac{n\pi y}{b}\,dy$ and then A_n is given by $A_n\sinh\frac{n\pi a}{b}+B_n\cosh\frac{n\pi a}{b}=\frac{2}{b}\int_0^b[f(y)-u_1-(u_2-u_1)\frac{y}{b}]\sin\frac{n\pi y}{b}\,dy$. Of course, to complete the evaluation we need to specify $p(y)$ and $f(y)$. **6.** Using $+\kappa^2$, in the first part of the solution, your solution should be identical to the solution, given above, to Exercise 4 but with $p(y)$ changed to u_3 and $f(y)$ changed to u_4.
10. (a) $u(x,y)=10-\frac{40}{\pi}\sum_{1,3,\dots}^\infty\frac{1}{n}\sin\frac{n\pi y}{2}e^{-n\pi x/2}$ **13.** $u(x,y)=20+6x+\sum_1^\infty\sin\frac{n\pi x}{5}(A_ne^{n\pi y/5}+B_ne^{-n\pi y/5})$, where $A_n+B_n=\frac{2}{5}\int_0^5[f(x)-20-6x]\sin\frac{n\pi x}{5}\,dx$. The latter is one equation in the two unknowns A_n and B_n, so we

can choose A_n *arbitrarily* and use the equation to solve for B_n. And if $A_n \neq 0$ then the $A_n e^{n\pi y/5}$ term is unbounded on the semi-infinite strip. **15.** (a) $u(x, y)$ is given by (15.2), where $U(x, y) = -\frac{fa}{2}x + \sum_1^\infty \sin\frac{n\pi x}{a}(A_n \cosh\frac{n\pi y}{a} + B_n \sinh\frac{n\pi y}{a})$; $A_n = \frac{f}{a}\int_0^a (ax - x^2)\sin\frac{n\pi x}{a}\,dx$ and $B_n = \frac{1-\cosh(n\pi b/a)}{\sinh(n\pi b/a)}A_n$

19. (a) $u(x, y) = (A + Bx)(C + Dy) + (E\cosh\kappa x + F\sinh\kappa x)(G\cosh\kappa y + H\sinh\kappa y)$, so $u(x, 0) = 0$ gives $C = G = 0$ so $u(x, y) = (P + Qx)y + (R\cosh\kappa x + S\sinh\kappa x)\sin\kappa y$. Then $u(x, 3) + 5u_y(x, 3) = 0$ gives $P = Q = 0$ and $\sin 3\kappa + 5\kappa\cos 3\kappa = 0$ with R and S remaining arbitrary. Thus, $u(x, y) = \sum_1^\infty(R_n\cosh\kappa_n x + S_n\sinh\kappa_n x)\sin\kappa_n y$, where κ_n's are the roots of $\tan\kappa = -5\kappa$. Finally, $u(0, y) = 0$ gives $R_n = 0$ and $u(4, y) = 100$ gives $S_n\sinh 4\kappa_n = \frac{\langle 100,\sin\kappa_n y\rangle}{\langle\sin\kappa_n y,\sin\kappa_n y\rangle}$ so $S_n = \frac{400}{\sinh 4\kappa_n}\frac{1-\cos 3\kappa_n}{6\kappa_n - \sin 6\kappa_n}$. [NOTE: The Sturm–Liouville problem is $Y'' + \kappa^2 Y = 0$ on $(0, 3)$ with $Y(0) = 0$ and $Y(3) + 5Y'(3) = 0$, so the inner product weight function is 1.] The *Maple* fsolve command gives $\kappa_1 = 0.6266$, $\kappa_2 = 1.6119$, $\kappa_3 = 2.6432$, $\kappa_4 = 3.6833$, $\kappa_5 = 4.7265$ so the first five terms of the series solution are $u(2, 1) = 18.6150 + 1.4048 + 0.0650 - 0.0056 - 0.0012 + \cdots = 20.078$. [Drawing the domain and boundary conditions, this value seems reasonable. If, on the other hand, we obtained $u(2, 1) = 83$ or 2.7, say, we would surely understand the calculation to be erroneous.]

Section 20.3

2. (a) $u(r, \theta) = 100\theta/\pi + \sum_1^\infty(A_n r^n + B_n r^{-n})\sin n\theta$, where $A_n = \frac{200}{n\pi}\frac{2^{-n}+(-1)^{n+1}}{2^{-n}-2^n}$ and $B_n = \frac{200}{n\pi}\frac{(-1)^n-2^n}{2^{-n}-2^n}$
(c) $u(r, \theta) = 100\theta/\pi + \frac{200}{\pi}\sum_1^\infty\frac{1}{n}\frac{(-1)^n}{2^n+2^{-n}}(r^n + r^{-n})\sin n\theta$ (e) $u(r, \theta) = 100\frac{\ln r}{\ln 2}$
(g) $u(r, \theta) = 100 - \frac{400}{\pi}\sum_{1,3,\dots}^\infty\frac{\sin(n\pi/2)}{n}\left(\frac{r}{3}\right)^{n/3}\cos\frac{n\theta}{3}$ NOTE: Applying the condition $u(3, \theta) = 0$ last gives $-100 = \sum_{1,3,\dots}^\infty A_n 3^{n/3}\cos\frac{n\theta}{3}$. Re-expressing the $\cos(n\theta/3)$ as $\cos\frac{n\pi\theta}{3\pi}$ we can identify "$2L$" $= 3\pi$, so "L" $= 3\pi/2$ in the quarter-range cosine formulas. **3.** (a) $u(r, \theta) = 50 + 20r\cos\theta$ **4.** Boundedness and the conditions at $\theta = 0$ and $\theta = 2\pi$ give $u(r, \theta) = A + \sum_1^\infty B_n b^{n/2}\cos\frac{n\theta}{2}$. Finally, $u(r, \theta) = 50 + 50\sin\theta = A + \sum_1^\infty B_n b^{n/2}\cos\frac{n\theta}{2}$ gives (by the half-range cosine formulas) A and B_n. The result is $u(r, \theta) = 50 + \frac{400}{\pi}\sum_{1,3,\dots}^\infty\frac{1}{4-n^2}\left(\frac{r}{b}\right)^{n/2}\cos\frac{n\theta}{2}$. No, the solution does not depend on the material. The unsteady diffusion equation $\alpha^2\nabla^2 u = u_t$ does contain the diffusivity α^2 of the material but in steady state $u_t = 0$ and the α^2 cancels out. **16.** (a) $u(r, z) = (A + B\ln r)(E + Fz) + [CJ_0(\kappa r) + DY_0(\kappa r)](Ge^{\kappa z} + He^{-\kappa z})$. Boundedness gives $u(r, z) = P + QJ_0(\kappa r)e^{-\kappa z}$, and $u(b, z) = 0$ gives $P = 0$ and $J_0(\kappa b) = 0$ or $\kappa = z_n/b \equiv \kappa_n$ $(n = 1, 2, \dots)$, where z_n is the nth positive root of $J_0(x) = 0$. Finally, $u(r, z) = \sum_1^\infty Q_n J_0(\kappa_n r)e^{-\kappa_n z}$ where $Q_n = \frac{2}{b^2[J_1(\kappa_n b)]^2}\int_0^b f(r)J_0(\kappa_n r)r\,dr$. (b) $u(r, z) = (A + B\ln r)(E + Fz) + [CI_0(\kappa r) + DK_0(\kappa r)](G\cos\kappa z + H\sin\kappa z)$. Boundedness as $r \to 0$ gives $B = D = 0$ and boundedness as $z \to \pm\infty$ gives $F = 0$, so $u(r, z) = P + I_0(\kappa r)(Q\cos\kappa z + R\sin\kappa z)$. We can superimpose any number of product solutions for different κ's so let us anticipate the Fourier series expansion of $u(b, z)$ and write $u(r, z) = P + \sum_1^\infty I_0\left(\frac{n\pi r}{L}\right)\left(Q_n\cos\frac{n\pi z}{L} + R_n\sin\frac{n\pi z}{L}\right)$. Then $u(b, z) = P + \sum_1^\infty I_0\left(\frac{n\pi b}{L}\right)\left(Q_n\cos\frac{n\pi z}{L} + R_n\sin\frac{n\pi z}{L}\right)$ gives (from the formulas for the Fourier series of a periodic function, with period $2L$) P, Q_n, and R_n. The result is $u(r, z) = 50 + \frac{200}{\pi}\sum_{1,3,\dots}^\infty\frac{1}{n}\frac{I_0(n\pi r/L)}{I_0(n\pi b/L)}\sin\frac{n\pi z}{L}$. **19.** (a) $u(\rho, \phi) = \sum_0^\infty A_n\rho^n P_n(\cos\phi)$, $u(\rho, \pi/2) = 0 = \sum_0^\infty A_n\rho^n P_n(0)$ so we need $A_n P_n(0) = 0$ for each $n = 0, 1, 2, \dots$. But $P_n(0) = 0$ for $n = 1, 3, \dots$ so we learn that $A_0, A_2, A_4, \dots$ are zero while $A_1, A_3, \dots$ remain arbitrary. Thus far, then, $u(\rho, \phi) = \sum_{1,3,\dots}^\infty A_n\rho^n P_n(\cos\phi)$. Finally, $u(c, \phi) = 100 = \sum_{1,3,\dots}^\infty A_n c^n P_n(\mu)$ so $A_n c^n = \frac{\langle 100,P_n(\mu)\rangle}{\langle P_n(\mu),P_n(\mu)\rangle} = \frac{\int_0^1 100P_n(\mu)\,d\mu}{\int_0^1 P_n^2(\mu)\,d\mu} = 100(2n + 1)\int_0^1 P_n(\mu)\,d\mu$. Evaluating these integrals gives $A_1 = 300/(2c)$, $A_3 = -700/(8c^3)$, $A_5 = 1100/(16c^5)$, $A_7 = -1500[5/(128c^7)]$, $A_9 = 1900[7/(256c^9)], \dots$, so $u(\rho, \mu) = 150\left(\frac{\rho}{c}\right)P_1(\mu) - \frac{175}{2}\left(\frac{\rho}{c}\right)^3 P_3(\mu) + \frac{275}{4}\left(\frac{\rho}{c}\right)^5 P_5(\mu) - \frac{1875}{32}\left(\frac{\rho}{c}\right)^7 P_7(\mu) + \frac{3325}{64}\left(\frac{\rho}{c}\right)^9 P_9(\mu) - \cdots$. As a partial check let us compute u on the z axis (so $\phi = 0$, $\mu = 1$) at $\rho = c/2$. Since $P_n(1) = 1$, we have $u(c/2, \mu = 1) = 75 - 10.94 + 2.15 - 0.46 + 0.10 - \cdots \approx 65.85$ which looks reasonable since the point is roughly midway between the flat bottom (on which $u = 0$) and the hemispherical top (on which $u = 100$), and the surface area of the hemisphere is twice that of the flat bottom. (b) $u(\rho, \phi) = \sum_0^\infty A_n\rho^n P_n(\cos\phi)$. On $\phi = \pi/2$ the unit outward normal is $\hat{\mathbf{n}} = \hat{\mathbf{e}}_\phi$ so $u_n = \nabla u\cdot\hat{\mathbf{n}} = \nabla u\cdot\hat{\mathbf{e}}_\phi = \frac{1}{\rho}\frac{\partial u}{\partial\phi}$. Thus, $u_n = 0$ on $\phi = \pi/2$ gives $u_\phi = 0$ there, so $u_\phi(\rho, \pi/2) = 0 = \sum_0^\infty A_n\rho^n P_n'(\cos\frac{\pi}{2})(-\sin\frac{\pi}{2}) = -\sum_0^\infty A_n P_n'(0)\rho^n$. Since

$P_n'(0) = 0$ for even n's and $P_n'(0) \neq 0$ for odd n's, it follows that $A_1 = A_3 = \cdots = 0$ with $A_0, A_2, A_4, \ldots$ remaining arbitrary, so $u(\rho, \phi) = \sum_{0,2,\ldots}^{\infty} A_n \rho^n P_n(\cos\phi)$. Finally, $u(c, \phi) = 100 = \sum_{0,2,\ldots}^{\infty} A_n c^n P_n(\mu)$ gives $A_n = 100 \frac{2n+1}{c^n} \int_0^1 P_n(\mu)\, d\mu$. Integration gives $A_0 = 100$, $A_2 = A_4 = \cdots = 0$ so the solution is simply the leading term, $u(\rho, \phi) = 100$, which does indeed satisfy the PDE and boundary conditions.

Section 20.4

1. (a) $u(x, y) = (100y/\pi) \int_0^{\infty} d\xi / [(\xi - x)^2 + y^2] = 50 - \frac{100}{\pi} \tan^{-1}\left(-\frac{x}{y}\right)$ where $\tan^{-1}()$ is to be taken as the "principal value," which is an odd function of its argument and which varies continuously from $-\pi/2$ at $() = -\infty$, through 0 at $() = 0$, to $+\pi/2$ at $() = +\infty$. Check: If $x > 0$ and $y \to 0$ through positive values then $u \to 50 - (100/\pi)\tan^{-1}(-\infty) = 50 + 50 = 100$, and if $x < 0$ and $y \to 0$ through positive values then $u \to 50 - (100/\pi)\tan^{-1}(\infty) = 0$, in agreement with the boundary conditions. NOTE: This problem is solved most easily using separation of variables in polar coordinates (the region being $0 < r < \infty$, $0 < \theta < \pi$) for in the solution form $u(r, \theta) = (A + B \ln r)(C + D\theta) + (Er^{\kappa} + Fr^{-\kappa})(G \cos \kappa\theta + H \sin \kappa\theta)$ boundedness as $r \to 0$ and as $r \to \infty$ gives $B = E = F = 0$ so $u = P + Q\theta$. Then $u(r, 0) = 100 = P$ and $u(r, \pi) = 0 = P + \pi Q$ give $u(r, \theta) = 100(1 - \frac{\theta}{\pi})$. If we write the latter as $u(r, \theta) = 100(1 - \frac{1}{\pi}\tan^{-1}\frac{y}{x})$ it doesn't look the same as our earlier solution but it must be remembered that $\tan^{-1}$ in the first solution is the branch of $\tan^{-1}$ lying between $-\pi/2$ and $\pi/2$, whereas $\tan^{-1}$ in the second solution is the branch lying between 0 and π (because $0 < \theta < \pi$).

Section 20.5

1. (a) $\alpha = \beta = \gamma = \delta = 1$ except at c, where $\alpha = \sqrt{12} - 3$ and $\delta = 3/4$; at e, where $\alpha = \sqrt{8} - 2$; and at g, where $\delta = 3/4$. Results: $u_a = 14.77$, $u_b = 18.24$, $u_c = 19.76$, $u_d = 13.12$, $u_e = 17.97$, $u_f = 14.58$ (d) Partial results: $u_a = 37.02$, $u_b = 65.83$, $u_c = 64.89$ (g) Partial results: $u_a = 48.26$, $u_b = 57.45$, $u_c = 45.71$. NOTE: At a we use the average value $U_W = (0 + 50)/2 = 25$. Similarly, at e we use $U_S = 25$. **2.** (b) The solution is symmetric about the lines $x = 0$, $y = 0$, $y = x$, and $y = -x$, so the unknowns reduce to $u(0, 0.8)$, $u(0.2, 0.8)$, $u(0.4, 0.8)$, $u(0.6, 0.8)$, and $u(0.8, 0.8)$, which values are found to be 50.30, 50.60, 52.08, 57.74, and 78.87, respectively. NOTE: From a sketch of the region, the grid, and the boundary conditions, these results look correct. **3.** (b) At a, $h = 0.5$, $\alpha = \beta = \gamma = \delta = 1$; at b, $h = 1$, $\alpha = 1$, $\beta = \gamma = \delta = 0.5$; at c, $h = 1$, $\alpha = \gamma = 1$, $\beta = \delta = 0.5$; and so on. $u_d = 5.051$, $u_c = 0.511$, $u_b = 0.060$, $u_a = 0.016$, $u_e = 0.004$, $u_g = 0.001$, $u_h = u_e$, $u_i = u_a$, $u_j = u_b$, $u_k = u_c$, $u_l = u_d$. Observe that the boundary condition input $u = 50$ at the ends are "felt," to any appreciable degree, only within a couple of end-widths of those ends; u_c is already only 0.511, and $u_b, u_c, \ldots, u_g$ are even smaller. This numerical result further illustrates the idea stated in Comment 3 of Example 1 (Section 20.2). **4.** (a) Partial results: Denoting the centerline points $(0.25, 0.5), (0.5, 0.5), \ldots, (2.5, 0.5)$ as a, b, c, d, e, f, respectively, the values obtained there are 53.13, 26.17, 12.40, 5.48, 1.11, 0.08. (b) By comparison, the exact values at those points are 54.47, 26.10, 12.03, 5.50, 1.14, 0.05. **6.** (a) The Taylor expansions of $u(x, y)$ about x_j, y_k, first with a step h to the right and then with a step h to the left, are

$$u(x_{j+1}, y_k) = u(x_j, y_k) + u_x(x_j, y_k)h + \frac{1}{2}u_{xx}(x_j, y_k)h^2 + \frac{1}{6}u_{xxx}(x_j, y_k)h^3 + \frac{1}{24}u_{xxxx}(x_j, y_k)h^4 + \cdots,$$
$$u(x_{j-1}, y_k) = u(x_j, y_k) - u_x(x_j, y_k)h + \frac{1}{2}u_{xx}(x_j, y_k)h^2 - \frac{1}{6}u_{xxx}(x_j, y_k)h^3 + \frac{1}{24}u_{xxxx}(x_j, y_k)h^4 - \cdots,$$

Addition gives $u_{xx}(x_j, y_k) = [u(x_{j+1}, y_k) - 2u(x_j, y_k) + u(x_{j-1}, y_k)]/h^2$ plus terms of order h^2 and higher. Similarly for $u_{yy}(x_j, y_k)$. When we drop those higher-order terms, in deriving (5), we incur a truncation error that is of order $O(h^2)$. (c) $h = 1/4$ gives $u_{\text{center}} = 21.339$, $h = 1/6$ gives $u_{\text{center}} = 20.570$, and the exact solution is $u_{\text{center}} = 19.927$. Then $19.927 - 21.339 = C\,(1/4)^p$ and $19.927 - 20.570 = C\,(1/6)^p$, which give $p = 1.94$.

Chapter 21

Section 21.2

9. (a) $2 - 11i$ (e) $-\frac{26}{125} - \frac{18}{125}i$ **11.** (a) $|z_1 + z_2| = |6 + 2i| = \sqrt{40} = 6.32$, $|z_1| + |z_2| = \sqrt{13} + \sqrt{17} = 7.73$

Section 21.3

2. (a) $2 < u < 3, 1 < v < 2$ (f) $\frac{u^2}{4} - 1 < v < 1 - \frac{u^2}{4}, -2 < u < 0$ **9.** (a) $-e^2$ (d) $\sin 3 \cosh \pi + i \cos 3 \sinh \pi$ (g) $\frac{2 \cosh 1 \sin 1}{\cosh 2 - \cos 2} + i \frac{2 \cos 1 \sinh 1}{\cosh 2 - \cos 2}$
16. (a) $\int_0^\infty e^{-x} \sin \omega x \, dx = \text{Im} \int_0^\infty e^{(i\omega - 1)x} \, dx = \text{Im} \, e^{(i\omega-1)x}/(i\omega - 1|_{x=0}^{x=\infty} = \text{Im} \, [1/(1 - i\omega)] = \omega/(1 + \omega^2)$
18. (a) $x_p(t) = \text{Im} \, \frac{F_0 e^{i\omega t}}{(k - m\omega^2) + i\omega c} = \text{Im} \, \frac{F_0 e^{i\omega t}}{(k - m\omega^2) + i\omega c} \frac{(k - m\omega^2) - i\omega c}{(k - m\omega^2) - i\omega c} = F_0 \frac{(k - m\omega^2) \sin \omega t - \omega c \cos \omega t}{(k - m\omega^2)^2 + \omega^2 c^2}$
(d) $x_p(t) = -\frac{10}{13}(5 \sin 5t + \cos 5t)$

Section 21.4

1. (a) $r = 3$, $\theta_0 = -\pi/2$ rad (f) $r = 2\sqrt{37}$, $\theta_0 = -1.406$ rad **4.** (a) $(-1 + i)^{10} = 32e^{i3\pi/2}$ (polar) $= -32i$ (Cartesian); $(-1 + i)^{20} = 1024e^{15\pi i}$ (polar) $= -1024$ (Cartesian) **5.** (a) $i^{1/2} = e^{i\pi/4}, e^{i5\pi/4}$; $i^{1/5} = e^{i\pi/10}, e^{i\pi/2}, e^{i9\pi/10}, e^{i13\pi/10}, e^{i17\pi/10}$ **6.** (a) $\log(-2) = \ln 2 + (2k + 1)\pi i$ $(k = 0, \pm 1, \pm 2, \ldots)$ **8.** (a) $(2i)^{2/3} = \sqrt[3]{4}e^{\pi i/3}, \sqrt[3]{4}e^{\pi i}, \sqrt[3]{4}e^{5\pi i/3}$; $(2i)^{3/2} = \sqrt{8}e^{3\pi i/4}, \sqrt{8}e^{7\pi i/4}$; $(2i)^\pi = e^{\pi \ln 2} e^{i(1+4k)\pi^2/2}$ $(k = 0, \pm 1, \pm 2, \ldots)$
9. (a) $(2i)^i = e^{-(\pi/2 + 2k\pi)}[\cos(\ln 2) + i \sin(\ln 2)]$; $(2i)^{1-i} = e^{\ln 2 + \pi/2 + 2k\pi}[\cos\left(\frac{\pi}{2} + 2k\pi - \ln 2\right)$
$+ i \sin\left(\frac{\pi}{2} + 2k\pi - \ln 2\right)]$ $(k = 0, \pm 1, \pm 2, \ldots)$ **11.** (a) $\log(-3i) = \ln 3 - i\frac{\pi}{2}$, $\sqrt{-3i} = \sqrt{\frac{3}{2}} - \sqrt{\frac{3}{2}}\, i$

Section 21.5

10. (a) $f'(z) = -\sin z$ **11.** (d) $f'(z) = -(2z + 3i)/(z^2 + 3iz - 2)^2$ for all $z \neq -i, -2i$ so $f(z)$ is analytic for all $z \neq -i, -2i$ **15.** (a) Harmonic for all z; $v(x, y) = e^x \sin y + c$; $f(z) = e^z +$ constant

Chapter 22

Section 22.2

5. (a) e^z analytic at $z = 0$ and $\frac{d}{dz}e^z = e^z \neq 0$ there, so yes, conformal at $z = 0$ (d) iz^2 analytic at $z = 0$ but $\frac{d}{dz}(iz^2) = 2iz = 0$ there, so no, not conformal at $z = 0$

Section 22.3

9. (a) Let $z_1 = i$, $z_2 = 0$, $z_3 = -i$ and $w_1 = 2i$, $w_2 = 2$, $w_3 = -2i$. Then (6.1) gives $w = \frac{2z+2}{-z+1}$. Note that this mapping must send the half plane $x < 0$ (rather than $x > 0$) into the interior $|w| < 2$ because as we walk along the boundary curve $x = 0$ from z_1 to z_2 to z_3 the region is on our right, so when we walk from w_1 to w_2 to w_3 the image region will again be our right. To obtain a second such mapping we could change z_1 to $5i$, say, and keep z_2, z_3, w_1, w_2, w_3 unchanged. (b) See the answer to part (a). To map onto the exterior $|w| > 2$ we can keep the z_j's unchanged and reverse the sequence of w_j's: $w_1 = -2i$, $w_2 = 2$, $w_3 = 2i$. **10.** (a) $z = 0 \rightarrow w = i$, $z = 1 \rightarrow w = 1 + i$, $z = \infty \rightarrow w = 1$ so the image is the interior of the circle through those three w points. To see that it is the interior and not the exterior we can use the idea presented in Exercise 8 or we can simply check one point: e.g., $z = i$ is outside of $\mathcal{D}$, so its image $w = \infty$ must be outside of $\mathcal{D}'$. (e) The image is the interior of a "crescent" with vertices at $w = 1$ and $w = i$. One arc of the crescent passes through $w = 1 + i$ and the other passes through $w = 2 + i$. **11.** (a) The image $\mathcal{D}'$ is bounded by a straight line from $0 + i\infty$ down to i, then a semicircle from i to $(1 + i)/2$ to 0, then a straight line from 0 to $0 - i\infty$ and is the region to the right of that boundary. (d) $\mathcal{D}'$ is the 90° sector with vertex at $w = 2$, with one edge running from 2 to $2 - i\infty$ and the other edge running from 2 to $-\infty$. **14.** (a) The mapping $w = (z-4)/(z+4)$ sends $z = -4$ to $w = \infty$ and $z = 4$ to $w = 0$. The image $\mathcal{D}'$ is the wedge with vertex at $w = 0$ with one edge from $w = 0$ to infinity through $w = (3+4i)/5$ and with the other edge from $w = 0$ to infinity through $w = (3-4i)/5$. Then $\Psi = A\phi + B$. $\Psi = 0$ on $\phi = -\alpha$ and $\Psi = 100$ on $\phi = +\alpha$ (where $\alpha = \tan^{-1}\frac{4}{3}$) gives $\Psi = 50(1 + \frac{\phi}{\alpha})$. But $\phi = \tan^{-1}(v/u) = \tan^{-1}[8y/(x^2+y^2-16)]$ because $u = (x^2+y^2-16)/[(x+4)^2+y^2]$ and $v = 8y/[(x+4)^2+y^2]$, so $\psi(x,(x,y) = 50\left(1 + \frac{\tan^{-1}[8y/(x^2+y^2-16)]}{\tan^{-1}(4/3)}\right)$, where $\tan^{-1}$'s are between $-\pi/2$ and $\pi/2$. Let us check as follows: At $x = 0$ and $y = 10$ we have $\psi(0,10) = 50(1 + \frac{0.761}{0.927}) = 91.0$, which seems reasonable. Further, as we approach infinity along any ray $y = mx$, $\psi \to 50$, which is also correct. (b) In entry 4, let $x_2 = 2$ and let $x_1 \to +\infty$. Then $a \sim (2x_1 + x_1\sqrt{3})/x_1 \to 2 + \sqrt{3}$ and $R \sim (2x_1 - x_1\sqrt{3})/x_1 \to 2 - \sqrt{3}$. $\Psi = A + B\ln\rho$ and the boundary conditions give $\Psi = 20 + 10\frac{\ln\rho}{\ln R}$ or $20 + 5\frac{\ln\rho^2}{\ln R}$. We can find $u(x,y)$ and $v(x,y)$ and then put $\rho^2 = u^2(x,y) + v^2(x,y)$ but it is simpler to write $\rho^2 = |w|^2 = \left|\frac{z-a}{az-1}\right|^2 = \left|\frac{(x-a)+iy}{(ax-1)+iay}\right|^2 = \frac{|(x-a)+iy|^2}{|(ax-1)+iay|^2}$, so $\psi(x,y) = 20 + \frac{5}{\ln R}\ln\frac{(x-a)^2+y^2}{(ax-1)^2+a^2y^2}$, where $R = 2-\sqrt{3}$ and $a = 2+\sqrt{3}$. As a check, let us compute $\psi(1.5,0) = 25.49$, which looks good. [Do you see why $\psi(1.5,0)$ should be just be slightly greater than 25?] (c) $w = \frac{z+3}{z-3}$ will map $\mathcal{D}$ into the strip $-1 < u < 0$.

Section 22.4

4. (a) $\mathcal{D}'$ is the upper half plane and $\Psi = 0$ on the entire u axis except on $0 < u < 1$, where $\Psi = 20$. Then $\Psi(u,v) = \frac{v}{\pi}\int_0^1 \frac{20\,du'}{(u'-u)^2+v^2} = \frac{20}{\pi}\left(\tan^{-1}\frac{u}{v} - \tan^{-1}\frac{u-1}{v}\right)$. But $w = u + iv = e^z = e^x(\cos y + i\sin y)$ so $u = e^x\cos y$, $v = e^x\sin y$, and $\psi(x,y) = \frac{20}{\pi}\left[\tan^{-1}(\cot y) - \tan^{-1}\left(\frac{e^x\cos y - 1}{e^x\sin y}\right)\right]$ where the $\tan^{-1}$'s are between $-\pi/2$ and $+\pi/2$. As a check, observe that on the midline of the strip $\psi(x,\pi/2) = \frac{20}{\pi}[\tan^{-1}0 - \tan^{-1}(-e^{-x})] = \frac{20}{\pi}\tan^{-1}e^{-x}$ does tend to 10 as $x \to -\infty$ (as it should) and to 0 as $x \to +\infty$ (as it should). (c) (12) in Section 20.4 gives $\Psi(u,v) = \frac{35}{2} - \frac{35}{\pi}\tan^{-1}(\frac{u+4}{v})$ so $\psi(x,y) = \frac{35}{2} - \frac{35}{\pi}\tan^{-1}(\frac{x^2-y^2+4}{2xy})$ where $-\pi/2 < \tan^{-1}() < \pi/2$. $\psi(2,2) = 12.33$ (which looks reasonable). (g) $\Psi(u,v) = 200 + \frac{400}{\pi}\tan^{-1}(\frac{u-1}{v})$ so $\psi(x,y) = 200 - \frac{400}{\pi}\tan^{-1}(\frac{\cos\pi x\cosh\pi y+1}{\sin\pi x\sinh\pi y})$ where $-\pi/2 < \tan^{-1}() < \pi/2$. Note that $\psi(0.5,y) \to 200$ as $y \to \infty$, as it should. Also, $\psi(0.5,0.1) = 39.4$.
(j) $\psi(x,y) = 30 - 5(x^2-y^2)/2$; $\psi(1,0) = 27.5$, $\psi(0.6,-0.4) = 29.5$

Section 22.5

3. (a) $\mathcal{D}'$ is the wedge $0 < \phi < 3\pi/2$ (i.e., the first three quadrants of the w plane); $\Psi(0,v) = \sin[\ln(-v)]$; $\Psi_N(u,0) = \frac{1}{|\exp(z)|}\psi_n(x,0) = (1/e^x)(-5) = -5/u$. Since $\mathcal{D}'$ is better suited to the polar coordinates ρ,ϕ, let us restate the mapped boundary conditions in terms of ρ,ϕ: $\Psi(\rho,3\pi/2) = \sin(\ln\rho)$ along $\phi = 3\pi/2$, and $\Psi_\phi(\rho,0) = 5$ along $\phi = 0$. To obtain the former, recall the directional derivative formula $\Psi_N = \hat{\mathbf{e}}_N\cdot\nabla\Psi$. On $\phi = 0$, $\hat{\mathbf{e}}_N$ is $-\hat{\mathbf{e}}_\phi$ so $\Psi_N = -\hat{\mathbf{e}}_\phi\cdot\nabla\Psi = -\frac{1}{\rho}\Psi_\phi(\rho,0)$ and setting this equal to $-5/\rho$ gives $\Psi_\phi(\rho,0) = 5$. (d) $\mathcal{D}'$ is the quarter of the unit disk $0 < \rho < 1$, $0 < \phi < \pi/2$. On $\phi = 0$, $\Psi = 5e^x = 5u$; on $\rho = 1$, $\Psi_N = 2y = 2\cos^{-1}u$; on $\phi = \pi/2$, $\Psi_N = 5/e^x = 5/v$. Or, in terms of polar variables ρ,ϕ, $\Psi(\rho,0) = 5\rho$, $\Psi_\rho(1,\phi) = 2\phi$, $\Psi_\phi(\rho,\pi/2) = 5$. (g) $\mathcal{D}'$ is the unit disk $\rho < 1$. The negative y axis maps to the lower semicircle so, in terms of the polar coordinates ρ,ϕ, $\Psi(1,\phi) = 50$ on $\pi < \phi < 2\pi$. The positive y axis maps to the upper semicircle so, since $y = 2v/[(1-u)^2+v^2] = v/(1-u) = \sin\phi/(1-\cos\phi)$, $\Psi(1,\phi) = 50\exp[-\sin\phi/(1-\cos\phi)]$ on $0 < \phi < \pi$. (j) $\mathcal{D}'$ is bounded below by $v = 0$ $(-\infty < u < 1)$ and on the right by $u = 1$ $(0 < v < \infty)$. The mapped boundary conditions are: $\Psi(u,0) = \sin\sqrt{-u}$ on $-\infty < u < 0$; $\Psi_v(u,0) = -1/(2\sqrt{u})$ on $0 < u < 1$; $\Psi(1,v) = 2\exp[-\sqrt{\sqrt{1+v^2} - 1}/\sqrt{2}]$ since $v = 2\sqrt{1+y^2}\,y$ gives $y = \sqrt{\sqrt{1+v^2} - 1}/\sqrt{2}$ by squaring and using the quadratic formula.

Section 22.6

2. (a) The mapping $\zeta = z^2$ gives the flow $W = U\zeta^2$ in the ζ plane and, hence, $w = U(z^2)^2 = Uz^4$ in the z plane. $u(x,y) = 4Ux(x^2-3y^2)$, $v(x,y) = 4Uy(y^2-3x^2)$, $\phi(x,y) = U[(x^2-y^2)^2 - 4x^2y^2]$, $\psi(x,y) = 4Uxy(x^2-y^2)$.

$u(0,0) = v(0,0) = 0$ so the corner is a stagnation point. (b) The mapping $\zeta = z^{2/3}$ gives the horizontal flow $dW/d\zeta = U$, $W = U\zeta$, $w = Uz^{2/3}$, $dw/dz = u - iv$ gives $u(r,\theta) = \frac{2U}{3r^{1/3}} \cos \frac{\theta}{3}$, $v(r,\theta) = \frac{2U}{3r^{1/3}} \sin \frac{\theta}{3}$. Rather than the origin being a stagnation point, $u \to \infty$ and $v \to \infty$ as $r \to 0$. (More generally, inside corners are stagnation points; outside corners are singular points.) **4.** (a) It is the flow given in Example 1, for any $U_0 > 0$.

Chapter 23

Section 23.2

1. (a) $\frac{2}{3}(1+i)$ (d) $-\pi i$ **3.** (a) Maximum $|z|$ on C is $\sqrt{8}$, and $L = \sqrt{5}$, so $|I| \leq (\sqrt{8})^5\sqrt{5} = 128\sqrt{10}$ (b) $|e^z| = |e^{x+iy}| = |e^x||e^{iy}| = e^x \leq e$ on C, and $L = 3\sqrt{2}$, so $|I| \leq 3\sqrt{2}e$ (c) $|e^{-z}| = e^{-x} \leq e^2$ on C, and $L = 3\sqrt{2}$, so $|I| \leq 3\sqrt{2}e^2$ (d) $\max|1/z| = 1/\min|z| = 1/(4-\sqrt{5})$, and $L = 8\pi$, so $|I| \leq 8\pi/(4-\sqrt{5})$

Section 23.3

4. (a) $\oint_{C_1} x\,dz = \int x d(e^{i\theta}) = i\int xe^{i\theta}\,d\theta = i\int_0^{2\pi} \cos\theta(\cos\theta + i\sin\theta)\,d\theta = \pi i$ (d) 0 by Cauchy's theorem since $\pm\sqrt{3}$ lie outside of C_3 (g) $-2\pi i/5$ (j) $\oint_{C_3} \frac{dz}{\sqrt{x^2+y^2}} = \int_{-1}^{1} \frac{i\,dy}{\sqrt{1+y^2}} + \int_1^{-1} \frac{dx}{\sqrt{x^2+1}} + \int_1^{-1} \frac{i\,dy}{\sqrt{1+y^2}} + \int_{-1}^{1} \frac{dx}{\sqrt{x^2+1}} = 0$ **6.** (a) Since z^{20} is analytic everywhere we can deform the contour to a straight line. Thus, $\int_C z^{20}\,dz = \int_0^1 x^{20}\,dx = 1/21$ **9.** (a) $I = \oint_C \frac{dz}{z-1} - \oint_C \frac{dz}{z} = 2\pi i - 2\pi i = 0$ (d) $I = -\oint_C \frac{dz}{z-1} + 2\oint_C \frac{dz}{z-2} = -2\pi i + 2(2\pi i) = 2\pi i$

Section 23.4

2. (a) $z^2/2$, $z^2/2+6$, $z^2/2+3-7i$ **3.** (a) $\int_0^i z\,dz = \frac{z^2}{2}\Big|_0^i = -\frac{1}{2}$ (d) $\int_i^0 \cos 3z\,dz = \frac{\sin 3z}{3}\Big|_i^0 = -\frac{\sin 3i}{3} = -\frac{i}{3}\sinh 3$ (g) $(e^{-9}-1)/2$ **4.** $(2n+\frac{1}{2})\pi i$

Section 23.5

1. (a) $2\pi i$ (d) $-4\pi i \sin 1$ (g) $\pi i(\sin 3 - 3\cos 3)$ (j) 0

Chapter 24

Section 24.2

5. (a) Ratio test: $L = 1/\sqrt{5} < 1$ so convergent (d) $c_n \to 1$ as $n \to \infty$ so (Theorem 24.2.2) divergent (g) $|c_n| = 1$ for all n, so c_n does not tend to 0 as $n \to \infty$; hence, divergent (Theorem 24.2.2) **6.** (a) $|z| < 1$ (d) $|z+i| < 1/e$ (g) $|z| < 1$ **8.** (a) The series converges in $|z-1| < 1$ so (Theorem 24.2.8) it is the Taylor series of the sum function, and it represents the sum function, in that disk. (d) Yes, it is the Taylor series of $1+z^3$ (about $z=0$) and it represents the function in $|z| < \infty$. **9.** (a) 1 (b) $\sqrt{10}$ **10.** (a) $\sqrt{4-\sqrt{6}}$, namely, the distance from the origin to the closer of the two roots of $z^2 - 2z + 3i + 1 = 0$ **11.** (a) $\sin z = z - \frac{1}{3!}z^3 + \frac{1}{5!}x^5 - \cdots, R = \infty$ (b) $\sin z = \sin a + (\cos a)t - \frac{\sin a}{2!}t^2 - \frac{\cos a}{3!}t^3 + \frac{\sin a}{4!}t^4 + \cdots, R = \infty$, where $t = z - a$, $\sin a = \sin(2-i) = \sin 2\cos i - \sin i\cos 2 = \sin 2\cosh 1 - i\sinh 1\cos 2$, $\cos a = \cos 2\cosh 1 + i\sin 2\sinh 1$ (d) Let $z^6 = w$. $e^{z^6} = e^w = 1 + w + \frac{1}{2!}w^2 + \frac{1}{3!}w^3 + \cdots = 1 + z^6 + \frac{1}{2!}z^{12} + \frac{1}{3!}z^{18} + \cdots, R = \infty$ **14.** (a) $f = z^{1/2}$, $f' = \frac{1}{2}z^{-1/2}$, $f'' = -\frac{1}{4}z^{-3/2}$, $f''' = \frac{3}{8}z^{-5/2}$, ... so $z^{1/2} = 1^{1/2} + \frac{1}{2}1^{-1/2}(z-1) - \frac{1}{4(2!)}1^{-3/2}(z-1)^2 + \frac{3}{8(3!)}1^{-5/2}(z-1)^3 + \cdots = 1^{1/2}\left[1 + \frac{1}{2}(z-1) - \frac{1}{4(2!)}(z-1)^2 + \frac{3}{8(3!)}(z-1)^3 - \cdots\right]$ (in $|z-1| < 1$), where, according to the branch cut, the $1^{1/2}$ factor is $+1$. (c) $z^{1/2} = i^{1/2} + \frac{1}{2}i^{-1/2}(z-1) - \frac{1}{4(2!)}i^{-3/2}(z-1)^2 + \frac{3}{8(3!)}i^{-5/2}(z-1)^3 - \cdots = i^{1/2}\left[1 + \frac{1}{2i}(z-1) - \frac{1}{4(2!)i^2}(z-1)^2 + \frac{3}{8(3!)i^3}(z-1)^3 - \cdots\right] = \frac{1+i}{\sqrt{2}}\left[1 - \frac{i}{2}(z-1) + \frac{1}{4(2!)}(z-1)^2\right.$

$+\frac{3i}{8(3!)}(z-1)^3 - \cdots\Big]$ in $|z-i|<1$ **16.** (a) $\tan z = z + \frac{1}{3}z^3 + \frac{2}{15}z^5 + \frac{17}{315}z^7 + \cdots$ in $|z| < \pi/2$ (d) $1 - z - z^2 + 5z^3 - 7z^4 - z^5 + 23z^6 - \cdots$ in $|z| < 1/\sqrt{3}$ (g) $\frac{1}{2} + \frac{1}{4}z + \frac{1}{8}z^2 + \frac{1}{48}z^3 - \frac{1}{96}z^4 - \frac{13}{960}z^5 - \cdots$. To determine the region of convergence, we seek the root of $2 - \sin z = 0$ that is closest to the origin. We could write $2 - \sin(x+iy) = 0$, put in $\sin(x+iy) = \sin x \cos iy + \sin iy \cos x = \sin x \cosh y + i \sinh y \cos x$, equate real and imaginary parts to 0 and solve the resulting two equations for x and y. But it is simpler to write $(e^{iz} - e^{-iz})/(2i) = 2$ or $t - 1/t = 4i$ where $t = e^{iz}$. Then $t^2 - 4it - 1 = 0$ gives $t = (2 \pm \sqrt{3})i$, so $iz = \log[(2 \pm \sqrt{3})i] = \ln(2 \pm \sqrt{3}) + i\left(\frac{\pi}{2} + 2n\pi\right)$ for $n = 0, \pm 1, \pm 2, \ldots$. Hence, $z = \frac{\pi}{2} + 2n\pi - i\ln(2 \pm \sqrt{3}) = \frac{\pi}{2} + 2n\pi \pm 1.317i$. The closest of these to the origin corresponds to $n = 0$ and gives the radius of convergence as $\sqrt{(\pi/2)^2 + (1.317)^2} = 2.05$.

Section 24.3

4. (a) $\frac{1}{z} = \sum_0^\infty \frac{(-i)^n}{(z-i)^{n+1}}$ (d) $\frac{1}{e^z - 1} = \frac{1}{z} - \frac{1}{2} + \frac{1}{12}z + \cdots$ (h) $\frac{1}{z^2} = \sum_0^\infty \frac{(n+1)i^n}{(z+1)^{n+2}}$ **5.** (a) $\sin\frac{1}{z} = \sum_0^\infty \frac{(-1)^n}{(2n+1)!}\frac{1}{z^{2n+1}}$ in $0 < |z| < \infty$ (b) Taylor series in $|z+2| < 2$: $\frac{1}{z} = -\frac{1}{2}\sum_0^\infty \left(\frac{z+2}{2}\right)^n$. Laurent series in $2 < |z+2| < \infty$: $\frac{1}{z} = \sum_0^\infty \frac{2^n}{(z+2)^{n+1}}$ **6.** $f(z) = \frac{1}{z^2}\left(1 + \frac{1}{z} + \frac{1}{z^2} + \cdots\right) = \frac{1}{z^2}\frac{1}{1-1/z} = \frac{1}{z^2 - z}$ so $f(2i) = -\frac{1}{5} + \frac{i}{10}$.

Section 24.4

2. (a) First-order zeros at $z = 0$, $z = 1$ (d) First-order zero at $z = 0$, second-order zeros at $z = (2n+1)\frac{\pi}{2}$ for $n = 0, \pm 1 \pm 2, \ldots$ **3.** (a) Second-order pole at $z = 0$ (d) First-order pole at $z = -2$ (g) Second-order pole at $z = 0$, third-order poles at $z = n\pi$ for $n = \pm 1, \pm 2, \ldots$ (j) Clearly analytic for all $z \neq 0$. For $z = 0$ expand $\sinh t$ in a Taylor series in t, $\sinh t = t + \frac{1}{3!}t^3 + \frac{1}{5!}t^5 + \cdots$ for $|t| < \infty$, so $\sinh\frac{1}{z} = \frac{1}{z} + \frac{1}{3!}\frac{1}{z^3} + \frac{1}{5!}\frac{1}{z^5} + \cdots$ in $|1/z| < \infty$, namely, in the annulus $0 < |z| < \infty$. Essential singularity at $z = 0$. **4.** (a) No, essential singularity there (d) Yes (g) No, essential singularity there (j) Yes **5.** (a) First-order pole. The key is to realize that the series does not converge in $0 < |z| < R$ for some R, but in $1 < |z| < \infty$. Thus, we cannot use the series to classify the singularity (if any) at $z = 0$. Rather, proceed as in Exercise 6 of Section 24.3.

Section 24.5

1. (a) 0 (d) $12\pi i$ (e) πi **2.** (a) $\pi\sqrt{2}/(4a^3)$ (d) $\pi/4$ (g) $3\pi/(4e^2)$ **3.** (a) $\pi/2$ (d) $\pi/4$ (g) $5\pi/16$ (j) $\pi/\sqrt{2}$ **6.** (a) $\pi/\sin(\pi a)$ (c) $3\pi/(4\sqrt{2})$ (d) 0 **9.** (a) $2\pi/(3\sqrt{3})$

Index