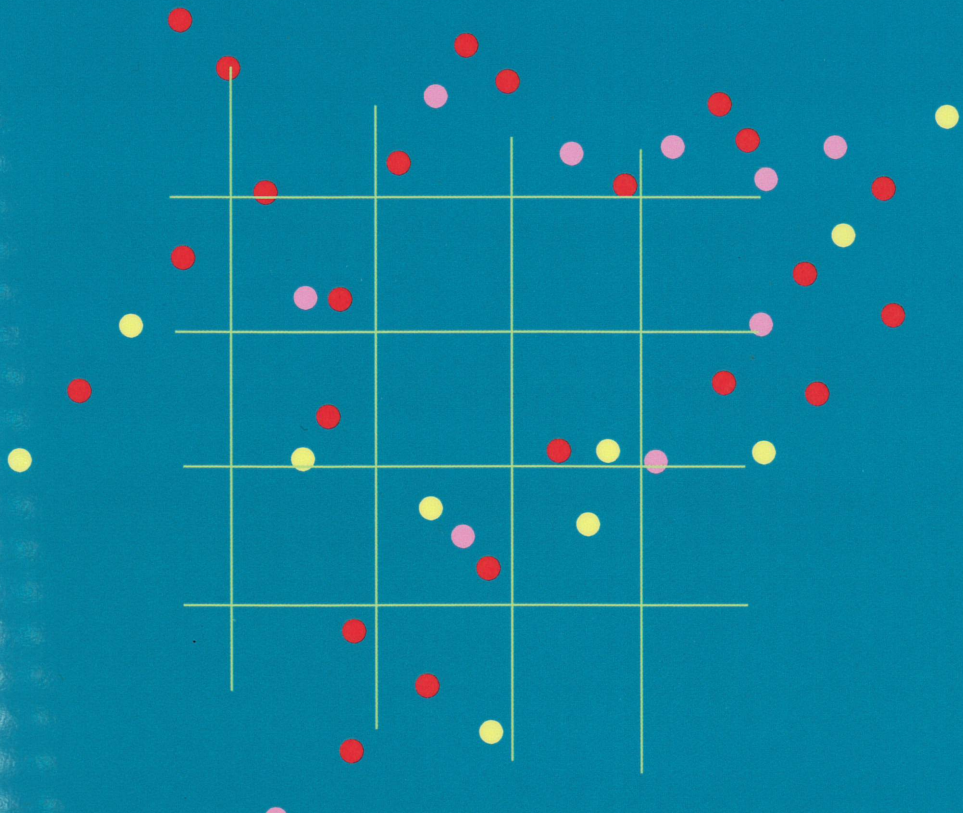
DIGITAL FILTERS

THIRD EDITION



R. W. HAMMING

DIGITAL FILTERS

Third Edition

R. W. HAMMING
Bell Laboratories

DOVER PUBLICATIONS, INC.
Mineola, New York

Copyright

Copyright © 1989 Lucent Technologies
All rights reserved.

Bibliographical Note

This Dover edition, first published in 1998, is an unabridged and slightly corrected republication of the work first published by Prentice-Hall, Inc., Englewood Cliffs, New Jersey in 1989. This work was written while the author was with Bell Laboratories. This Dover edition is published by special arrangement with Bell Laboratories, Lucent Technologies.

Library of Congress Cataloging-in-Publication Data

Hamming, R. W. (Richard Wesley), 1915–
Digital filters / R. W. Hamming.
p. cm.
Originally published: 3rd ed. Englewood Cliffs, NJ : Prentice-Hall, 1989.
Includes bibliographical references and index.
ISBN-13: 978-0-486-65088-3
ISBN-10: 0-486-65088-X
1. Digital filters (Mathematics) I. Title.
QA297.H26 1998
621.3815'324—dc21 97-51370
CIP

Manufactured in the United States by Courier Corporation
65088X08 2013
www.doverpublications.com

Contents

PREFACE TO THE THIRD EDITION xi

1 INTRODUCTION 1

1.1 What Is a Digital Filter? *1*
1.2 Why Should We Care About Digital Filters? *7*
1.3 How Shall We Treat the Subject? *10*
1.4 General-Purpose Versus Special-Purpose Computers *10*
1.5 Assumed Statistical Background *11*
1.6 The Distribution of a Statistic *15*
1.7 Noise Amplification in a Filter *17*
1.8 Geometric Progressions *19*

2 THE FREQUENCY APPROACH 21

2.1 Introduction *21*
2.2 Aliasing *22*
2.3 The Idea of an Eigenfunction *25*
2.4 Invariance Under Translation *28*
2.5 Linear Systems *31*
2.6 The Eigenfunctions of Equally Spaced Sampling *33*
2.7 Summary *34*

3 SOME CLASSICAL APPLICATIONS 36

3.1 Introduction *36*
3.2 Least-Squares Fitting of Polynomials *37*
3.3 Least-Squares Quadratics and Quartics *42*
3.4 Modified Least Squares *47*
3.5 Differences and Derivatives *50*
3.6 More on Smoothing: Decibles *54*
3.7 Missing Data and Interpolation *56*
3.8 A Class of Nonrecursive Smoothing Filters *59*
3.9 An Example of How a Filter Works *63*
3.10 Integration: Recursive Filters *66*
3.11 Summary *70*

4 FOURIER SERIES: CONTINUOUS CASE 71

4.1 Need for the Theory *71*
4.2 Orthogonality *72*
4.3 Formal Expansions *75*
4.4 Odd and Even Functions *82*
4.5 Fourier Series and Least Squares *86*
4.6 Class of Functions and Rate of Convergence *88*
4.7 Convergence at a Point of Continuity *90*
4.8 Convergence at a Point of Discontinuity *94*
4.9 The Complex Fourier Series 95
4.10 The Phase Form of a Fourier Series 99

5 WINDOWS 102

5.1 Introduction *102*
5.2 Generating New Fourier Series: The Convolution Theorems *103*
5.3 The Gibbs Phenomenon *107*
5.4 Lanczos Smoothing: The Sigma Factors *109*
5.5 The Gibbs Phenomenon Again *112*
5.6 Modified Fourier Series *115*
5.7 The von Hann Window: The Raised Cosine Window *116*

5.8 Hamming Window: Raised Cosine with a Platform *118*
5.9 Review of Windows *121*

6 DESIGN OF NONRECURSIVE FILTERS **124**

6.1 Introduction *124*
6.2 A Low-Pass Filter Design *127*
6.3 Continuous Design Methods: A Review *130*
6.4 A Differentiation Filter 133
6.5 Testing the Differentiating Filter on Data *138*
6.6 New Filters from Old Ones: Sharpening a Filter *140*
6.7 Bandpass Differentiators *147*
6.8 Midpoint Formulas *147*

7 SMOOTH NONRECURSIVE FILTERS **150**

7.1 Objections to Ripples in a Transfer Function *150*
7.2 Smooth Filters *153*
7.3 Transforming to the Fourier Series *157*
7.4 Polynomial Processing in General *159*
7.5 The Design of a Smooth Filter *160*
7.6 Smooth Bandpass Filters *162*

8 THE FOURIER INTEGRAL AND THE SAMPLING THEOREM **164**

8.1 Introduction *164*
8.2 Summary of Results *165*
8.3 The Sampling Theorem *166*
8.4 The Fourier Integral *169*
8.5 Some Transform Pairs *170*
8.6 Band-Limited Functions and the Sampling Theorem *173*
8.7 The Convolution Theorem *176*
8.8 The Effect of a Finite Sample Size *177*
8.9 Windows *179*
8.10 The Uncertainty Principle *181*

9 KAISER WINDOWS AND OPTIMIZATION 185

9.1 Windows *185*
9.2 Review of Gibbs Phenomenon and the Rectangular Window *187*
9.3 The Kaiser Window: I_0-sinh Window *189*
9.4 Derivation of the Kaiser Formulas *194*
9.5 Design of a Bandpass Filter *195*
9.6 Review of Kaiser Window Filter Design *196*
9.7 The Same Differentiator Again *198*
9.8 A Particular Case of Differentiation *199*
9.9 Optimizing a Design *203*
9.10 A Crude Method of Optimizing *204*

10 THE FINITE FOURIER SERIES 208

10.1 Introduction *208*
10.2 Orthogonality *208*
10.3 Relationship Between the Discrete and Continuous Expansions *212*
10.4 The Fast Fourier Transform *214*
10.5 Cosine Expansions *216*
10.6 Another Method of Design 217
10.7 Padding Out Zeros *217*

11 THE SPECTRUM 219

11.1 Review *219*
11.2 Finite Sample Effects *220*
11.3 Aliasing *221*
11.4 Computing the Spectrum *221*
11.5 Nonharmonic Frequencies *223*
11.6 Removal of the Mean *224*
11.7 The Phase Spectrum *227*
11.8 Summary *228*

12 RECURSIVE FILTERS 230

12.1 Why Recursive Filters? *230*
12.2 Linear Differential Equation Theory *233*
12.3 Linear Difference Equations *238*

12.4 Reduction to Simpler Form *240*
12.5 Stability and the *Z* Transformation *244*
12.6 Butterworth Filters *246*
12.7 A Simple Case of Butterworth Filter Design *250*
12.8 Removing the Phase: Two-Way Filters *252*

13 CHEBYSHEV APPROXIMATION AND CHEBYSHEV FILTERS **253**

13.1 Introduction *253*
13.2 Chebyshev Polynomials *254*
13.3 The Chebyshev Criterion *256*
13.4 Chebyshev Filters *258*
13.5 Chebyshev Filters, Type 1 *258*
13.6 Chebyshev Filters, Type 2 *263*
13.7 Elliptic Filters *264*
13.8 Leveling an Error Curve *264*
13.9 A Chebyshev Identity *265*
13.10 An Example of the Design of an Integrator *267*
13.11 Phase-Free Recursive Filters *270*
13.12 The Transient *270*

14 MISCELLANEOUS **272**

14.1 Types of Filter Design *272*
14.2 Finite Arithmetic Effects *273*
14.3 Recursive Versus Nonrecursive Filters *275*
14.4 Direct Modeling *276*
14.5 Decimation *277*
14.6 Time-Varying Filters *277*
14.7 References *278*

INDEX **281**

Preface to the Third Edition

This edition retains the same basic approach of the earlier editions of stressing fundamentals; however, some changes have been made to reflect the fact that increasingly often a digital filter course is the first course in electrical engineering and the field of signal processing. To meet these needs two main changes have been made: (1) the inclusion of more material on the z-transform, which is often used in later courses (though the constant use of the formalism tends to obscure the ideas behind the manipulations), and (2) the inclusion of more examples and exercises. There are, of course, many minor changes to clarify and adapt the material to current uses.

In the years since I wrote the first edition I have become increasingly convinced of the need for a very elementary treatment of the subject of digital filters. The need for an elementary introduction comes from the fact that many of the people who most need the knowledge are not mathematically sophisticated and do not have an elaborate electrical engineering background. Thus this book assumes *only* a knowledge of the calculus and a smattering of statistics (which is reviewed in the text). It does not assume any electrical engineering background knowledge. Actually, experience seems to show that a prior knowledge of the corresponding theory of analog filters often causes more harm than good! Digital filtering is not simply converting from analog to digital filters; it is a fundamentally different way of thinking about the topic of signal processing, and many of the ideas and limitations of the analog method have no counterpart in the digital form.

The subject of digital filters is the natural introduction to the broad, fundamental field of signal processing. The power and basic simplicity of digital signal processing over the older analog is so great that whenever possible we are converting present analog systems to an equivalent digital

form. But much more important, digital signaling allows fundamentally new things to be done easily. The availability of modern integrated-circuit chips, as well as micro- and minicomputers, has greatly expanded the application of digital filters.

Digital signals occur in many places. The telephone company is rapidly converting to the use of digital signals to represent the human voice. Even radio, television, and hi-fi sound systems are moving toward the all digital methods since they provide such superior fidelity and freedom from noise, as well as much more flexible signal processing. The space shots use digital signaling to transmit the information from the planets back to Earth, including the extremely detailed pictures (which were often processed digitally here on Earth to extract further information and to form alternate views of what was originally captured by the cameras in space). Most records of laboratory experiments are now recorded in digital form, from isolated measurements using a digital voltmeter to the automatic recording of entire sets of functions via a digital computer. Thus these signals are immediately ready for digital signal processing to extract the message that the experiment was designed to reveal. Economic data, from stock market prices and averages to the Cost of Living Index of the Bureau of Labor Statistics, occur only in digital form.

Digital filtering includes the processes of smoothing, predicting, differentiating, integrating, separation of signals, and removal of noise from a signal. Thus many people who do such things are actually using digital filters without realizing that they are; being unacquainted with the theory, they neither understand what they have done nor the possibilities of what they might have done. Computer people very often find themselves involved in filtering signals when they have had no appropriate training at all. Their needs are especially catered to in this book.

Because the same ideas arise in many fields there are many cross connections between the fields that can be exploited. Unfortunately each field seems to go its own way (while reinventing the wheel) and to develop its own jargon for exactly the same ideas that are used elsewhere. One goal of this revision is to expose and reduce this elaborate jargon equivalence from the various fields of application and to provide a unified approach to the whole field. We will adopt the simplest, most easily understood words to describe what is going on and exhibit lists of the equivalent words from related fields. We will also use only the simplest, most direct mathematical tools and shun fancy mathematics whenever possible.

This book concentrates on linear signal processing; the main exceptions are the examination of roundoff effects and a brief mention of Kalman filters, which adapt themselves to the signal they are receiving.

The fundamental tool of digital filtering is the frequency approach, which is based on the use of sines and cosines rather than on the use of polynomials (as is conventional in many fields such as numerical analysis

and much of statistics). The frequency approach, which leads to the spectrum, has been the principal method of opening the black boxes of nature. Examples run from the early study of the structure of the atom (using spectral lines as the observations) through quantum mechanics (which arose from the study of the spectrum of black-body radiation) to the modern methods of studying a system (for purposes of modeling and control) via the spectrum of the output as it is related to the input.

There appears to be a deep emotional resistance to the frequency approach. And even electrical engineers who use it daily often have only a slight understanding of *why* they are using the eigenfunction approach and the role of the eigenvalues. In numerical analysis there is almost complete antipathy to the frequency approach, while in statistics there is a great fondness for polynomials (without ever examining the question of which set of functions is appropriate). This book shows clearly why the sines and cosines are the natural, the proper, the characteristic functions to use in many situations. It also approaches cautiously the usual traumatic experience (for most people) of going from the real sines and cosines to the complex exponentials with the mysterious $\sqrt{-1}$; their greater convenience in use eventually compensates for the initial troubles and provides more insight.

The text includes an accurate (but not excessively rigorous) introduction to the necessary mathematics. In each case the formal mathematics is postponed until the need for it is clearly seen. We are interested in presenting the *ideas* of the field and will generally not give the "best" methods for designing very complex filters; in an elementary course it is proper to give elementary, broadly applicable design methods, and then show how these can be refined to meet a very wide range of design criteria. Because it is an elementary text, references to advanced papers and books are of little use to the reader. Instead we refer to a few standard texts where more advanced material and references can be found. The references to these books are indicated in the text by [L,p], where L is the book label given at the end of this book, and p is the page(s) where it can be found. References [IEEE-1 and 2] give a complete bibliography for most topics that arise.

There is a deliberate repetition in the presentation of the material. Experience shows that the learner often becomes so involved in the immediate details of designing a filter that where and how the topic fits into the whole plan is lost. Furthermore, confusion often arises when the same ideas and mathematical tools are used in seemingly very different situations. It is also true that filters are designed to process data, but experience shows that the display of large sets of data that have been processed communicates very little to the beginner. Thus such plots are seldom given, even though the learner needs to be reminded that the ultimate test of a filter is how well it processes a signal, not how elegant the derivation is.

As always an author is deeply indebted to others, in this case to his many colleagues at Bell Laboratories. Special mention should go to Pro-

fessor J. W. Tukey (of Princeton University) and to J. F. Kaiser, who first taught him most of what is presented here. Thanks are also due to Roger Pinkham and the many students of the short courses who used the first two editions; their questions and reactions have been important in many places of this revision. They have also strengthened the author's belief in the basic rightness of giving as simple an approach as possible and of keeping rigorous mathematics in its proper place. Finally, thanks are due to the Naval Postgraduate School for providing an atmosphere suitable for thinking deeply about the problems of teaching.

R. W. Hamming

1

Introduction

1.1 WHAT IS A DIGITAL FILTER?

In our current technical society we often measure a continuously varying quantity. Some examples include blood pressure, earthquake displacements, voltage from a voice signal in a telephone conversation, brightness of a star, population of a city, waves falling on a beach, and the probability of death. All these measurements vary with time; we regard them as functions of time: $u(t)$ in mathematical notation. And we may be concerned with blood pressure measurements from moment to moment or from year to year. Furthermore, we may be concerned with functions whose independent variable is not time, for example the number of particles that decay in a physics experiment as a function of the energy of the emitted particle. Usually these variables can be regarded as varying continuously (analog signals) even if, as with the population of a city, a bacterial colony, or the number of particles in the physics experiment, the number being measured must change by unit amounts.

For technical reasons, instead of the signal $u(t)$, we usually record *equally spaced samples* u_n of the function $u(t)$. The famous *sampling theorem*, which will be discussed in Chapter 8, gives the conditions on the signal that justify this sampling process. Moreover, when the samples are taken they are not recorded with infinite precision but are rounded off (sometimes chopped off) to comparatively few digits (see Figure 1.1-1). This procedure is often called *quantizing* the samples. It is these quantized samples that are available for the processing that we do. We do the processing in order to understand what the function samples u_n reveal about the underlying phenomena that gave rise to the observations, and digital filters are the main processing tool.

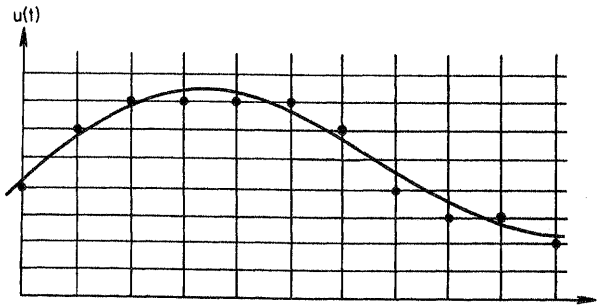


FIGURE 1.1-1 SAMPLING AND QUANTIZATION OF A SIGNAL

It is necessary to emphasize that the samples are *assumed* to be equally spaced; any error or *noise* is in the measurements u_n. Fortunately, this assumption is approximately true in most applications.

Suppose that the sequence of numbers $\{u_n\}$ is such a set of equally spaced measurements of some quantity $u(t)$, where n is an integer and t is a continuous variable. Typically, t represents time, but not necessarily so. We are using the notation $u_n = u(n)$. The simplest kinds of filters are the *nonrecursive filters*; they are defined by the linear formula

$$y_n = \sum_{k=-\infty}^{\infty} c_k u_{n-k} \tag{1.1-1}$$

The coefficients c_k are the constants of the filter, the u_{n-k} are the input data, and the y_n are the outputs. Figure 1.1-2 shows how this formula is computed. Imagine two strips of paper. On the first strip, written one below the other, are the data values u_{n-k}. On the second strip, with the values written in the *reverse direction* (from bottom to top), are the filter coefficients c_k. The zero subscript of one is opposite the n subscript value of the other (either way). The output y_n is the sum of all the products $c_k u_{n-k}$. Having computed one value, one strip, say the coefficient strip, is moved one space down, and the new set of products is computed to give the new output y_{n+1}. Each output is the result of adding all the products formed from the proper displacement between the two zero-subscripted terms. In the computer, of course, it is the data that is "run past" the coefficient array $\{c_k\}$.

This process is basic and is called a *convolution* of the data with the coefficients. It does not matter which strip is written in the reverse order; the result is the same. So the convolution of u_n with the coefficients c_k is the same as the convolution of the coefficients c_k with the data u_n.

In practice, the number of products we can handle must be finite. It is usual to assume that the length of the run of nonzero coefficients c_k is much shorter than is the run of data y_n. Once in a while it is useful to regard the

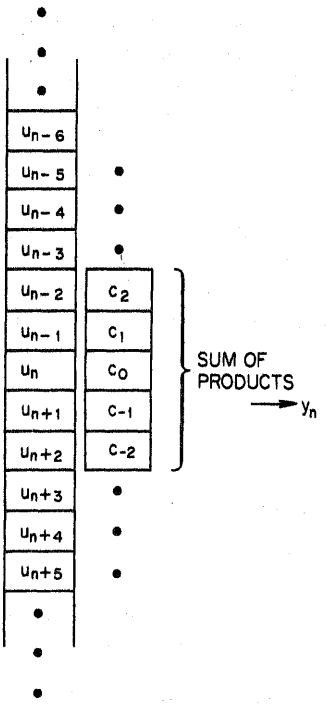


FIGURE 1.1-2 A NONRECURSIVE DIGITAL FILTER

c_k coefficients as part of an infinite array with many zero coefficients, but it is usually preferable to think of the array $\{c_k\}$ as being finite and to ignore the zero terms beyond the end of the array. Equation (1.1-1) becomes, therefore,

$$y_n = \sum_{k=-N}^{N} c_k u_{n-k} \tag{1.1-2}$$

Thus the second strip (of coefficients c_k) in Figure 1.1-2 is comparatively shorter than is the first strip (of data u_n).

Various special cases of this formula occur frequently and should be familiar to most readers. Indeed, such formulas are so commonplace that a book could be devoted to their listing. In the case of five nonzero coefficients c_k, where all the coefficients that are not zero have the same value, we have the familiar smoothing by 5*s* formula (derived in Section 3.2)

$$y_n = \tfrac{1}{5}(u_{n-2} + u_{n-1} + u_n + u_{n+1} + u_{n+2}) \tag{1.1-3}$$

Another example is the least-squares smoothing formula derived by passing a least-squares cubic through five equally spaced values u_n and using the value of the cubic at the midpoint as the smoothed value. The formula for this smoothed value (which will be derived in Section 3.3) is

$$y_n = \tfrac{1}{35}(-3u_{n-2} + 12u_{n-1} + 17u_n + 12u_{n+1} - 3u_{n+2}) \quad (1.1\text{-}4)$$

Many other formulas, such as those for predicting stock market prices, as well as other time series, also are nonrecursive filters.

Nonrecursive filters occur in many different fields and, as a result, have acquired many different names. Among the disguises are the following:

Finite impulse response filter

FIR filter

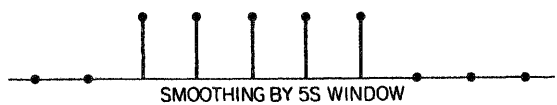
Transversal filter

Tapped delay line filter

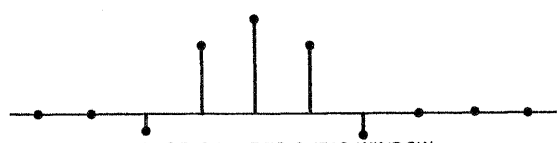
Moving average filter

We shall use the name *nonrecursive* as it is the simplest to understand from its name, and it contrasts with the name *recursive filter*, which we will soon introduce.

The concept of a *window* is perhaps the most confusing concept in the whole subject, so we now introduce it in these simple cases. We can think of the preceding formulas as if we were looking at the data u_{n-k} through a *window of coefficients* c_k (see Figure 1.1-3). As we slide the strip of coefficients along the data, we see the data in the form of the output y_n, which



SMOOTHING BY 5S WINDOW



LEAST-SQUARES CUBIC WINDOW

FIGURE 1.1-3 WINDOWS

is the running weighted average of the original data u_n. It is as if we saw the data through a translucent (not transparent) window where the window was tinted according to the coefficients c_k. In the smoothing by $5s$, all data values get through the translucent window with the same amount, $\frac{1}{5}$; in the second example they come through the window with varying weights. (Don't let any negative weights bother you, since we are merely using a manner of speaking when we use the words "translucent window.")

When we use not only data values to compute the output values y_n but also use other values of the output, we have a formula of the form

$$y_n = \sum_{-\infty}^{\infty} c_k u_{n-k} + \sum_{-\infty}^{\infty} d_k y_{n-k}$$

where both the c_k and the d_k are constants. In this case it is usual to limit the range of nonzero coefficients to current and past values of the data u_n and to only past values of the output y_n. Furthermore, again the number of products that can be computed in practice must be finite. Thus the formula is usually written in the form

$$y_n = \sum_{0}^{N} c_k u_{n-k} + \sum_{1}^{M} d_k y_{n-k} \tag{1.1-5}$$

where there may be some zero coefficients. These are called *recursive* filters (see Figure 1.1-4). Some equivalent names follow:

Infinite impulse response filter

IIR filter

Ladder filter

Lattice filter

Wave digital filter

Autoregressive moving average filter

ARMA filter

Autoregressive integrated moving average filter

ARIMA filter

We shall use the name *recursive* filter. A recursive digital filter is simply a linear difference equation with constant coefficients and nothing more; in practice it may be realized by a short program on a general purpose digital computer or by a special purpose integrated circuit chip.

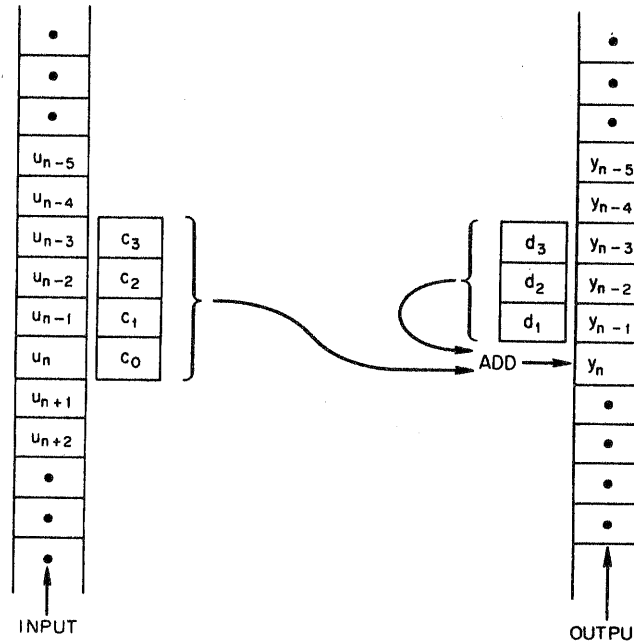


FIGURE 1.1-4 RECURSIVE DIGITAL FILTER

A familiar example (from the calculus) of a recursive filter is the trapezoid rule for integration

$$y_n = y_{n-1} + \tfrac{1}{2}[u_n + u_{n-1}] \tag{1.1-6}$$

It is immediately obvious that a recursive filter can, as it were, remember all the past data, since the y_{n-1} value on the right side of the equation enters into the computation of the new value y_n, and hence into the computation of y_{n+1}, y_{n+2}, and so on. In this way the initial condition for the integration is "remembered" throughout the entire estimation of the integral.

Other examples of a recursive digital filter are the exponential smoothing forecast

$$y_{n+1} = au_{n+1} + (1 - a)y_n \qquad (0 < a < 1)$$

and the trend indicator

$$T_n = c[u_n - u_{n-1}] + (1 - c)T_{n-1} \qquad (0 < c < 1)$$

As is customary, we have set aside recursive filters that use *future*

values, values beyond the currently computed value. If we used future values beyond the current y_n, we would have to solve a system of linear algebraic equations, and this is a lot of computing. At times it is worth it, but often we have only past computed values and the current value of the data. Filters that use only past and current values of the data are called *causal*, for if time is the independent variable, they do not react to future events but only past ones (causes).

It is worth a careful note, however, that more and more often all the data of an experiment is recorded on a magnetic tape or other storage medium *before* any data processing is done. In such cases the restriction to causal filters is plainly foolish. Future values are available! There are, of course, many situations in which the data must be reduced and used as they come in, and in such cases the restriction to causal filters is natural.

The student may wonder where we get the starting values of the y_n. Once well going they, of course, come from previous computations, but how to start? The custom of assuming that the missing values y are to be taken as zeros is very dubious. This assumption usually amounts to putting a sharp discontinuity into the function y_n, and since as noted previously the recursive filter remembers the past, it follows that these zero values continue to affect the computation for some time, if not indefinitely. It is evident in the simple example of the trapezoid integration that the needed starting value of y is the starting area, usually taken to be zero, but not necessarily so.

We have said it before, but it is necessary to say again that the coefficients c_k and d_k of the filter are assumed to be constants. Such filters are called *time-invariant filters* and are the filters most used in practice. Time-varying filters are occasionally useful and will be briefly touched upon in this text.

Finally, it should be realized that in practice all computing must be done with finite-length numbers. The process of quantization affects not only the input numbers, but it may affect all the internal (to the filter) arithmetic that is done. Consequently, there are roundoff errors in the final output numbers y_n. It is often convenient to *think* in terms of infinite precision arithmetic and perfect input data u_n; but in the end we must deal with reality. Furthermore, the details of the way we arrange to do the arithmetic can affect the accuracy of the output numbers. We will look at this topic more closely in the closing chapters.

1.2 WHY SHOULD WE CARE ABOUT DIGITAL FILTERS?

The word *filter* is derived from electrical engineering, where filters are used to transform electrical signals from one form to another, especially to eliminate (filter out) various frequencies in a signal. As we have already

seen, a digital filter is a linear combination of the input data u_n and possibly the output data y_n and includes many of the operations that we do when processing a signal.

For convenience we suppose that we sample at unit time and that we represent the nth sample of the signal as u_n. Thus we may think of blood pressure, a brain wave, the height of a wave on a beach, or a stock market price as a continuous signal that we have sampled (and quantized) at unit times in order to obtain our sequence of data u_n. In the stock market case we often take the integral over a period, say a week, and record only the total amount of stock sold each week, although the underlying idea of continuous variation of the price or whatever else is being measured (say the rate at which shares are traded) still exists. Given such a signal, we may want to differentiate, integrate, sum, difference, smooth, extrapolate, analyze for periodicity, or possibly remove the noise; all these, and many others, are linear operations. Therefore, in the digital form, the operations are *digital filters*.

Widespread use of mini- and microcomputers in science, medicine, and engineering has greatly increased the number of digital signals recorded and processed. Since we are already processing such data in a linear fashion, it is necessary to understand the alterations and distortions that these filters produce. Moreover, because digital transmission is so much more noise resistant than is analog signal transmission, a world dominated by digital transmission is rapidly approaching. Thus again we are impelled to study exactly what digital filters do, or can be designed to do, to various signals.

Applications of digital filters now greatly transcend those that arise in electrical engineering. As a result, it is necessary to redefine and remove some of the restrictions that were natural to electrical engineering at the time when digital filters were emerging from the classical electrical analog filters. The student should carefully note that we are sometimes making *different definitions than those that frequently occur in older electrical engineering texts*. It is necessary to do this because we have a larger view of the field of applications; we include applications to numerical analysis and statistics, for example, as well as to other fields.

Occasionally you read about *sampled data systems*. Here the signal is sampled, but the sampled value is *not* quantized. Such systems will not be considered in this book.

We will always assume that the samples u_n are unit spaced and begin at $t_0 = 0$. If they are not, it is easy to find the linear transformation that will make them so. Let the original data be at

$$t_n = t_0 + n\,\Delta t \qquad (n = 0, 1, 2, \ldots)$$

To find the transformation, we simply assume the form

$$t'_n = at_n + b$$

and impose the two conditions

$$t_0' = 0 = at_0 + b$$
$$t_1' = 1 = a(t_0 + \Delta t) + b$$

Solve (by subtraction)

$$1 = a\,\Delta t$$
$$b = -at_0$$

Therefore,

$$t_n' = \frac{1}{\Delta t}(t_n - t_0) \tag{1.2-1}$$

It is easy to see that this is the desired transformation.

The method of derivation should be learned rather than the result memorized.

Exercises

1.2-1 Find the standard transformation for the data: $t_0 = 10$, $t_1 = 12$, $t_2 = 14, \ldots,$ *Answer:* $t' = \frac{1}{2}(t_n - 10)$.

1.2-2 Find the transformation that moves the sample points $t_0 = 0.100$, $t_1 = 0.112$, $t_2 = 0.124, \ldots,$ to the standard form.

1.2-3 List ten sources, other than those in the text, of signals that might be filtered.

1.2-4 Write Simpson's integration formula as a recursive filter.

1.2-5 Compute the first five output values of the filter

$$y_n = ay_{n-1} + u_n$$

for the input $y_0 = 0$, $u_n = 1$ for all n.

1.2-6 Compute the successive values of the filter

$$y_n = ay_{n-1} + u_n$$

where $y_0 = 0$, $u_1 = 1$, and all other $u_n = 0$. Give the formula for y_n.

1.2-7 Compact disc (CD) recordings use 44.1 kH (kH = kiloHertz = 1000 times a second). Find the appropriate transformation.

1.3 HOW SHALL WE TREAT THE SUBJECT?

Much of the theory, both as to the design and use of digital filters, originated in the field of analog filters. If one is already familiar with the field, then it might be reasonable to build on this knowledge. Today, however, the average person who needs to know about digital filters has no such background, and so it is foolish to base the development on the analog approach. Consequently, we assume no such familiarity and will only mention the corresponding jargon when necessary.

The statistics field has also contributed extensively to the theory of digital filters. In particular, the subject of time series is closely related and has contributed its own elaborate and confusing jargon.

Textbooks in numerical analysis have many formulas that are linear combinations of equally spaced data, and thus such formulas are equivalent to digital filters. Since the elements of numerical analysis are now more widely known than those of other fields of application, we will select many of our examples from numerical analysis. Furthermore it is very often needed in practice.

The fundamental approach common to all the special fields is based on (1) the Fourier series, both discrete and continuous, and (2) the use of the Fourier integral. They are the mathematical tools for understanding and manipulating linear formulas, and we must take the time to develop them, for they are rarely taught outside of electrical engineering courses these days. However, we will avoid becoming too involved with mathematical rigor, which all too often tends to become rigor mortis. Nor do we develop all the mathematical theory before showing its use; instead, we regularly give applications of the theory just covered in order to show both its relevance and its use. In this way, we hope that much of the mathematics will become more obvious to the nonmathematically inclined.

1.4 GENERAL-PURPOSE VERSUS SPECIAL-PURPOSE COMPUTERS

Digital filtering is done using both special- and general-purpose digital computers. Even though numerical computations also use both types, most introductory textbooks on the subject deal only with computing done on general-purpose computers; similarly, most of the discussion in this book is confined to filtering done on general-purpose computers.

This remark should not be interpreted as meaning that the field of special-purpose computers is unimportant. Rather it is an indication that computation on a general-purpose computer is usually much less restrictive. Therefore, in a first presentation, we concentrate on the main ideas, while

ignoring the details of the particular computer being used. Special-purpose digital computers are rapidly increasing in importance, primarily because of the availability of inexpensive, large-scale integrated circuits, as well as the fact that many of the operations that we wish to perform are sometimes beyond the scope (in either an economic or a time sense, or both) of current (and foreseeable) general-purpose computers.

1.5 ASSUMED STATISTICAL BACKGROUND

We need to take a brief look at statistics. A set of measurements is called a *sample*. The word "sample" is used both for an individual measurement and for the set of measurements, even if they are repeated measurements of the same thing. This usage occurs because the statistician is thinking of an underlying *population* or *ensemble* of possible measurements and you have obtained one possible set of results (one realization). The statistician is concerned with the probability of obtaining the particular observed result and with the effects of repetitions of the experiment. The measurements of a sample may all be at a single point. For instance, the sample can be a number of measurements of the length of a particular wire. The measurements can also be scattered at various places in the range of a function, for example, the velocity of a boat at various times of a day.

Often a *model* for the distribution of the measurements must be found; we want to think about the ensemble from which we have drawn the particular sample.

To illustrate, if L is the measured length of the wire just mentioned, then we model these measurements by

$$P\{L \leq x\} = P(x)$$

For $P\{L \leq x\}$ read "probability that L is less than or equal to x." $P(x)$ is thus the probability that the measured length L is less than or equal to x; $P(x)$ is called the *cumulative distribution function for* L. In many situations $P(x)$ has a derivative $p(x)$; that is,

$$\frac{dP}{dx} = p(x) \quad \text{and} \quad P\{a < L \leq b\} = \int_a^b p(x)\,dx \tag{1.5-1}$$

Then $p(x)$ is called either the *density* or the *probability density for* L.

A common density that occurs in such situations as measuring the length of a piece of wire is the gaussian (or normal) distribution

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2} \qquad (-\infty < x < \infty) \tag{1.5-2}$$

where μ and σ are parameters whose values depend on the particular situation being modeled. See Figure 1.5-1(b).

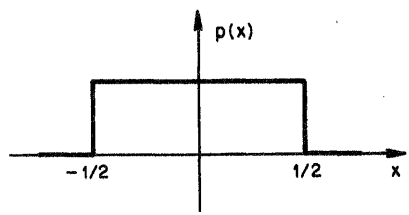
Another example of a model occurs in roundoff theory. It is reasonable to suppose that the roundoff error made when a number is quantized (rounded off) is "uniformly distributed" from $-\frac{1}{2}$ to $\frac{1}{2}$ in the last digit kept. Therefore,

$$p(x) \equiv \begin{cases} 1, & -\frac{1}{2} \le x \le \frac{1}{2} \\ 0, & |x| > \frac{1}{2} \end{cases} \tag{1.5-3}$$

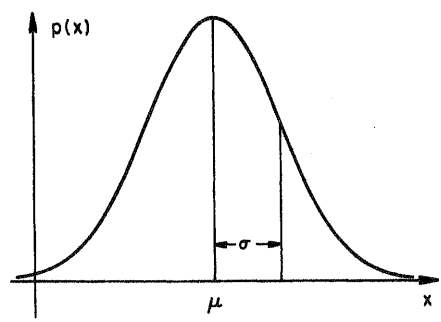
See Figure 1.5-1(a).

A commonly computed characteristic of a random quantity such as L or, alternatively, of a density $p(x)$ is the *average* or *expected value* (also called *mean value*). It is denoted Ave $\{L\}$ or $E\{L\}$ and is defined by

$$\text{Ave}\,\{L\} \equiv E\{L\} \equiv \int_{-\infty}^{\infty} xp(x)\,dx \tag{1.5-4}$$



(a) ROUNDOFF DISTRIBUTION



(b) GAUSSIAN DISTRIBUTION

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

FIGURE 1.5-1

You are weighting each x by its probability of occurring, $p(x)$, and combining them all together. For the roundoff example,

$$\int_{-\infty}^{\infty} xp(x)\,dx = \int_{-1/2}^{1/2} x\,dx = 0$$

Note that the averaging is over the ensemble $p(x)$.

For the gaussian example (1.5-2),

$$E\{L\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-1/2[(x-\mu)/\sigma]^2}\,dx$$

Replacing $(x - \mu)/\sigma$ by the normalized variable t we have

$$\begin{aligned} E\{L\} &= \sigma \int_{-\infty}^{\infty} te^{-(1/2)t^2} \frac{dt}{\sqrt{2\pi}} + \mu \int_{-\infty}^{\infty} e^{-(1/2)t^2} \frac{dt}{\sqrt{2\pi}} \\ &= 0 + \mu = \mu \end{aligned} \tag{1.5-5}$$

The last equation is true because the first integrand is odd and therefore the integral equals zero. The second integral is $P\{-\infty < L < \infty\} = 1$.

We can think of the expectation as an operator $E\{\ \}$ operating on a function. A moment's thought and it is obvious that the expected value of a constant is the same constant,

$$E\{a\} = a$$

Again, if x is the variable of the model and if a and b are constants, then

$$E\{ax + b\} = aE\{x\} + b$$

Other "typical values" besides the average are widely used. One is the *mode*, the most frequent value or the one with maximum probability density. Another is the *median*, the value exceeded by half the distribution. We will not use them in this book.

Another commonly computed characteristic of a random quantity or, alternatively, of its distribution is the *variance*. It is denoted Var $\{L\}$ if L is the random quantity, and is defined by

$$\text{Var}\,\{L\} = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)\,dx \tag{1.5-6}$$

where L has density $p(x)$ and $\mu = E\{L\}$. It is also denoted by the symbol σ^2. Note that the variance is always measured about the mean. In mechanics this same expression is known as the *moment of inertia*.

For the roundoff case (1.5-3), we have

$$\sigma^2 = \int_{-\infty}^{\infty} (x - 0)^2 p(x)\, dx = \int_{-1/2}^{1/2} x^2\, dx = \frac{1}{12} \tag{1.5-7}$$

and for the gaussian (1.5-2),

$$\sigma^2 = \text{Var}\,\{L\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 e^{-1/2[(x-\mu)/\sigma]^2}\, dx$$

If we set

$$x - \mu = \sigma t$$

we obtain

$$\text{Var}\,\{x\} = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2/2}\, dt$$

Integration by parts using

$$\begin{cases} te^{-t^2/2}\, dt = dV, & V = -e^{-t^2/2} \\ U = t, & dU = dt \end{cases}$$

gives

$$\text{Var}\,\{x\} = \frac{\sigma^2}{\sqrt{2\pi}} \left[-e^{-t^2/2} t \,\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-t^2/2}\, dt \right]$$

The integrated piece vanishes at both ends, and in the integral $\mu = 0$, $\sigma^2 = 1$. Therefore, since the integral equals $\sqrt{2\pi}$, we have

$$\text{Var}\,\{x\} = \frac{\sigma^2}{\sqrt{2\pi}} \sqrt{2\pi} = \sigma^2$$

This result shows why we adopted the peculiar form for writing the gaussian distribution: σ^2 is the variance and μ is the average of the gaussian (normal) distribution,

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \equiv N(\mu, \sigma^2)$$

It is clear that the variance, which is the sum of the squares of the

deviations of the distribution from its average value [weighted by the probability $p(x)$ of occurring], is closely related to the principle of least squares (which states that the *best fit* occurs when the sum of the squares of the errors is minimum). In both cases, it is the sum of the squares of the differences that is used. For the variance, it is the difference from the mean that is used; for a least-squares fit, it is the difference of the data from the approximate fit that is used.

Exercises

1.5-1 If the distribution for $p(x)$ is

$$p(x) = \begin{cases} 0, & x < 0 \\ ae^{-ax}, & x \geq 0 \end{cases} \qquad (a > 0)$$

show that $\mu = 1/a$ and $\sigma^2 = 1/a^2$.

1.5-2 If the distribution for $p(x)$ is

$$p(x) = \begin{cases} 1 - \dfrac{x}{2}, & 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

show that $\mu = \frac{2}{3}$, $\sigma^2 = \frac{2}{9}$.

1.5-3 Find the mean and variance of

$$p(x) = \begin{cases} \cos 2x, & -\dfrac{\pi}{4} \leq x \leq \dfrac{\pi}{4} \\ 0, & \text{otherwise} \end{cases}$$

1.5-4 For a well-balanced die (singular of dice), calculate the mean and variance of the value on the top face after a random toss. Do the same for a pair of dice.

1.6 THE DISTRIBUTION OF A STATISTIC

We now turn to what is probably the hardest concept for the beginner in statistics to master, the idea of the distribution of a statistic (such as the mean or the variance of a sample).

Suppose that we have made a set of measurements and that from the sample we have computed one or more statistics. For instance, we may have selected 1000 Americans at random from the entire population of around 200 million and measured their heights. From these heights we can compute the *sample average* $\bar{x}$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The *variance* s^2 *of the sample* is defined by

$$s^2 \equiv \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

For clarity, it is customary to use Greek letters for the statistics of the model and Latin letters for the corresponding statistics of the sample.

It is good to know these two numbers for the sample that we drew, but if we are to make much use of them, the question immediately arises: If we repeat the whole process again, using a different random sample of 1000 Americans, what could we reasonably expect to get for the average? In short, what is the distribution of the statistic called the "average"? Clearly, repetitions of the whole process of selecting the people, making the measurements, and computing the average will give us a distribution of values for the average $\bar{x}$ (and a distribution for the variance s^2).

In the roundoff example we had a unique model for the basic population from which the roundoffs were drawn, but in the gaussian example we must estimate the two unknown parameters of the population distribution μ and σ^2 from the sample statistics $\bar{x}$ and s^2. We can ask what relation these two sets of numbers have to each other. In textbooks on statistics it is proved that, for any distribution, the *average* of the sample is an unbiased estimate of the original population average. Similarly, the sample variance s^2 is an unbiased estimate of σ^2. Unbiased means that, on the average, your estimates are neither too high nor too low. (That is, the average of the statistic equals the value being estimated.)

TABLE 1.6-1 Relation of sample to population statistics

Sample	Population
$\text{Ave}(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$	μ
$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$	σ^2

If the sample is at all large ($n \geq 10$), then the central limit theorem shows that the statistic called the average has a distribution that is very close to a gaussian (normal) distribution

$$p(x) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-n(x-\bar{x})^2/2\sigma^2}$$

with parameters $\bar{x}$ and σ^2/n.

Exercise

1.6-1 For the set of measurements 10, 11, 10, 12, 9, 10, 7, 10, 10, 9, compute the mean and variance of the sample and estimate the corresponding population parameters.

1.7 NOISE AMPLIFICATION IN A FILTER

Suppose that we make some measurements. Let u_n be the "true" measurement with added noise ϵ_n whose expected value is zero. The condition of zero mean is

$$E\{\epsilon_n\} = 0$$

and implies that there is no *bias* in the measurements, only local, random errors. The averaging is, of course, over the ensemble of noise ϵ_n. Furthermore, let this noise ϵ_n have a variance σ^2. What is the corresponding noise in the output of a nonrecursive filter (assuming that the arithmetic we do does not increase the noise)? To compute this, we make the additional assumption (which is often, but not always, true) that in making the measurements $u_n + \epsilon_n$ the errors ϵ_n are *uncorrelated*. This assumption in mathematical notation is

$$E\{\epsilon_n\epsilon_m\} = \begin{cases} \sigma^2, & m = n \\ 0, & m \neq n \end{cases}$$

Again the averaging is over the ensemble of the noise.

A nonrecursive filter is defined by the formula

$$y_n = \sum_{k=-N}^{N} c_k(u_{n-k} + \epsilon_{n-k}) \tag{1.7-1}$$

The expected value is, therefore, since the E operation applies only to the ϵ_n and not to the c_k or the u_{n-k}, and $E\{\epsilon_n\} = 0$,

$$E\{y_n\} = \sum_{k=-N}^{N} c_k(u_{n-k} + E\{\epsilon_{n-k}\}) = \sum_{k=-N}^{N} c_k u_{n-k}$$

For the variance calculation, we begin with

$$E\left\{\left[\sum_{k=-N}^{N} c_k(u_{n-k} + \epsilon_{n-k}) - E(y_n)\right]^2\right\}$$

But this is (using different summation indices to keep things clear)

$$E\left\{\sum_{k=-N}^{N} c_k\epsilon_{n-k}\right\}^2 = E\left\{\left[\sum_{k=-N}^{N} c_k\epsilon_{n-k}\right]\left[\sum_{m=-N}^{N} c_m\epsilon_{n-m}\right]\right\}$$

Since $E\{\epsilon_n\} = 0$ and, for $m \neq n$, $E\{\epsilon_n\epsilon_m\} = 0$, multiplying out and applying the operator E to the ϵ_n leaves only the terms

$$\sum c_k^2 E\{\epsilon_k^2\} = \sum c_k^2\sigma^2 = \sigma^2 \sum_{k=-N}^{N} c_k^2 \tag{1.7-2}$$

Thus the sum of the squares of the coefficients of a filter measures the noise amplification of the filtering process. It is for this reason that the sum of the squares of the coefficients of a nonrecursive filter plays a significant role in the theory.

Exercises

1.7-1 Apply formula (1.7-2) using the roundoff noise model.

1.7-2 What is the noise amplification of the least-squares cubic (1.1-4) of Section 1.1?

1.7-3 What is the noise amplification of smoothing by 5*s*? *Answer:* $\sigma^2 = \frac{1}{5}$.

1.7-4 Show that the minimum noise amplification of a five-term nonrecursive filter, with the sum of the coefficients equal to 1, is the smoothing by 5*s*. *Hint:* Use Lagrange multipliers.

1.8 GEOMETRIC PROGRESSIONS

We need to review the topic of geometric progressions since it will occur often in the text.

A *geometric progression* is a sequence of n values

$$a, az, az^2, \ldots, az^{n-1} \tag{1.8-1}$$

where each is obtained from the previous one by multiplying by the constant z. The sum of all the terms is

$$S(n) = a(1 + z + z^2 + \cdots + z^{n-1}) \tag{1.8-2}$$

If we multiply this equation by z and then subtract it from the original we get

$$S(n) - zS(n) = a - az^n$$

$$S(n) = \frac{a(1 - z^n)}{1 - z} \tag{1.8-3}$$

The special case of symmetrically arranged terms beginning at $a = z^{-m}$ and going on for $2m + 1$ terms gives

$$z^{-m} + z^{-m+1} + \cdots + 1 + \cdots z^{m-1} + z^m$$

This sum we will label as $S(-m, m)$ and is, from (1.8-3),

$$S(-m, m) = \frac{z^{-m}(1 - z^{2m+1})}{1 - z} = \frac{z^{-m} - z^{m+1}}{1 - z}$$

Multiply the numerator and denominator by $-z^{-1/2}$

$$S(-m, m) = \frac{z^{m+1/2} - z^{-(m+1/2)}}{z^{1/2} - z^{-1/2}} \tag{1.8-4}$$

In this derivation z may be a real or complex number. Differentiating (1.8-2) and (1.8-3) with respect to z we get ($a = 1$)

$$1 + 2z + 3z^2 + \cdots + (n - 1)z^{n-2} = \frac{1 - nz^{n-1} + (n - 1)z^n}{(1 - z)^2}$$

Multiply by z to get the convenient formula

$$z + 2z^2 + 3z^3 + \cdots + (n-1)z^{n-1} = z\frac{[1 - nz^{n-1} + (n-1)z^n]}{(1-z)^2} \tag{1.8-5}$$

If $|z| < 1$ then letting $n \to \infty$ (1.8-3) becomes

$$S(\infty) = \frac{a}{1-z} \tag{1.8-6}$$

and (1.8-5) becomes

$$\sum_{k=0}^{\infty} kz^k = \frac{z}{(1-z)^2} \tag{1.8-7}$$

We often want $1/z$ in place of z. Then (1.8-3) becomes

$$a(1 + 1/z + 1/z^2 + \cdots + 1/z^{n-1}) = \frac{az(1 - z^{-n})}{z-1} \tag{1.8-8}$$

and (1.8-6) is

$$a\sum_{k=0}^{\infty} z^{-k} = \frac{az}{z-1} \tag{1.8-9}$$

while (1.8-7) is

$$\sum_{k=0}^{\infty} kz^{-k} = \frac{z}{(z-1)^2} \tag{1.8-10}$$

Similar formulas are easily found by similar methods.

Exercises 1.8

1.8-1 Discuss (1.8-4) when z is replaced by $1/z$.

1.8-2 From (1.8-10) find

$$\sum_{k=0}^{\infty} k^2 z^{-k}$$

1.8-3 In (1.8-5) replace z by $1/z$.

2

The Frequency Approach

2.1 INTRODUCTION

The purpose of this chapter is to show, for linear digital filter design, why and in what sense the use of sines and cosines of the independent variable t is preferable to the classical use of polynomials in t. The approximation of a function by a polynomial is generally emphasized in mathematics, statistics, and numerical analysis. For instance, in Newton's method for finding a zero of a function $g(t)$, the function is locally replaced by the tangent line, a linear equation in t. Again, a Taylor's expansion of a function expresses the function in powers of $t - t_0$. In statistics data is constantly being fitted by polynomials. In the trapezoid rule for integration the function is locally replaced by a straight line. It is natural, therefore, to suppose that in other fields polynomials are the proper functions to use when approximating a given function. Thus we are concerned in this chapter more with the psychological problem of undoing this earlier conditioning in favor of polynomials than with the logical problem of presenting the frequency approach.

Before doing this, however, we introduce in the next section the most important consequence of sampling a function at equally spaced points. This phenomenon, called *aliasing*, is a common experience for most people; but they are so accustomed to it that they are only vaguely aware of it.

We shall then show, in three different senses, that the sines and cosines are the proper functions for situations that are relevant to much of data processing on computers. To do so, it is necessary to introduce the concepts of *eigenfunctions* and *eigenvalues* and to show that the concept of the *transfer function* corresponds to the eigenvalues of the process.

Since the idea of *frequency* is clearly central to the frequency approach,

we must be careful to say what it means. Consider, for example, a rectangular wave (or any other shaped wave) that exactly repeats itself 10 times a second; we say that it has a *period* (cycle) of $T = \frac{1}{10}$ second and a *rotational frequency* of 10 hertz (cycles per second). Hertz is abbreviated Hz when used as a unit of measure. By a period of a function, we mean the *shortest* interval for which the function exactly repeats itself, and the *fundamental frequency* is the corresponding frequency.

The period T and the frequency f are reciprocals of each other. The *angular frequency* ω (in radians) is related to the rotational frequency f by

$$\omega = 2\pi f \tag{2.1-1}$$

The use of the angular measure ω is natural in calculus situations, and use of the rotational frequency f is natural in applications. This is the way that we will use them. This use of two different units occurs similarly in logs; in the theory part, as in the calculus, we use natural logs, while in practice we use the logs to the base 10.

The adjective *fundamental* is often dropped, but doing so can lead to confusion. For instance, in Section 4.3 we will decompose a rectangular wave into a sum of sines and cosines and then say that the original wave form has high frequencies in it. Thus confusion can arise concerning the frequency of the original wave form and the frequencies of the terms in the decomposition of the wave into a set of sinusoidal (periodic) functions.

2.2 ALIASING

The phenomenon of *aliasing*, which is basic to sampling data at equally spaced intervals, is not new to the reader who has watched Westerns either on television or in the movies. As the stagecoach wheels turn faster and faster, they appear to slow down and then to stop. If the increase in speed is great enough, they may seem to go backward, stop, and go forward a number of times. Any actual high rate of rotation of the wheels appears, as a result of the equally spaced *sampling of the pictures* in time, to be "aliased" into a low frequency of rotation. Figure 2.2-1 shows, symbolically, a wheel with four spokes rotating at different rates, and the human mind interprets what is seen as the smallest motion that accounts for the observations. At the first time that the wheel seems to stop, it will appear to have twice the normal number of spokes.

Another common application of this phenomenon of *aliasing due to sampling* occurs when a stroboscope is flashed at a rate close to that of a piece of rotating equipment. If the stroboscope flashes at a rate slightly *less* than the rate of rotation (or some multiple of it), then the flashes make the

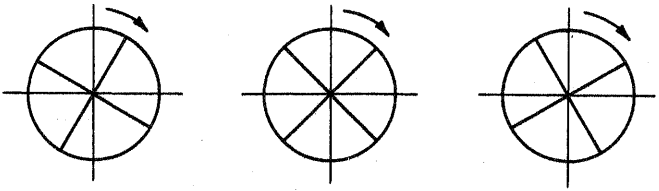


FIGURE 2.2-1

machine appear to the eye as if it were rotating slowly forward; the closer the rates, the slower the apparent rotation. Again we see that one frequency is aliased into another due to the process of taking equally spaced samples.

In the case of a sinusoid, we are not sampling a rotating wheel but are, in effect, sampling one component, either vertical or horizontal (or in any other direction for that matter). As a result, we can see the phenomenon of aliasing due to sampling at equal intervals in time (the independent variable) as a simple consequence of trigonometric identities.

Consider the sinusoid

$$u(t) = \cos[2\pi(m + a)t + \phi]$$

where m is an integer, positive or negative, a is the *positive* fractional part of the original rate of rotation, and ϕ is an arbitrary phase angle. Since we are sampling the sinusoid at integer values of t, the reduction of any angle by $2m\pi$ leaves the cosine with the same values at the sample points. Therefore, the sinusoid is equivalent (at the sample points $t_n = n$) to

$$\cos[2\pi at + \phi]$$

If $a > \frac{1}{2}$, we can remove another 2π, and [since $\cos x = \cos(-x)$]

$$\cos[2\pi(-1 + a)t + \phi] = \cos[2\pi(1 - a)t - \phi]$$

at the sample points. Thus we have shown, using only simple trigonometry, that *at the sample points* any sinusoid of arbitrary frequency is equivalent to a sinusoid with a frequency that lies between 0 and $\frac{1}{2}$, equivalent in the sense that the two sinusoids have the same numerical values at the sample points (see Figure 2.2-2). In a very real sense, the two frequencies are indistinguishable: a high frequency is aliased into (appears as) a low frequency

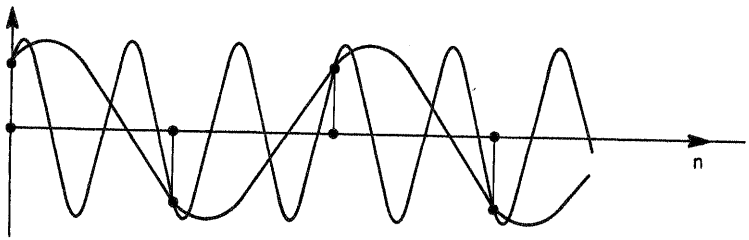


FIGURE 2.2-2 ALIASING

due solely to the sampling process. Only sinusoids with frequencies low enough so that at least two samples occur in each period are not aliased.

To illustrate, the function

$$u_n = \cos \frac{7\pi}{2} n \qquad (n = 0, \pm 1, \pm 2, \ldots)$$

has the same values as

$$u_n = \cos\left(\frac{7\pi}{2} - 4\pi\right)n = \cos\left(\frac{-\pi}{2} n\right) = \cos \frac{\pi}{2} n$$

at all the sample points, and hence the original function is aliased into the lower-frequency function.

Exercises

2.2-1 A machine rotates at 100 Hz. If a strobe light flashes at a rate of 99 per second, what is the apparent motion of the machine? (*Hint:* Take $\frac{1}{99}$ second as the unit of time for the sampling.) If the strobe flashes 101 times per second? 98 times?

2.2-2 Find the lowest aliased frequency of $\cos[8\pi n/3 + \pi/3]$. Of $\cos[13\pi n/3 + \pi/3]$.

2.2-3 Repeat the argument for sines in place of cosines. Note carefully the small differences between the cosines and sines.

2.2-4 What is the aliased frequency of $\cos 4\pi n$?

2.2-5 What is the highest frequency of a constant?

2.2-6 A compact disc recording uses 44,100 samples per second. What is the highest unaliased frequency? *Answer:* 22,050

2.3 THE IDEA OF AN EIGENFUNCTION

The word *eigenfunction* is a half-translation from the German of what in the older English texts was called characteristic function, proper function, or natural function.

For purposes of illustration *only* consider a special case of eigenfunctions, the multiplication of a square matrix $\mathbf{A} = (a_{i,j})$, of dimension N by N, by a vector $\mathbf{x}$, of dimension N by 1. The product is another vector $\mathbf{y}$, of dimension N by 1,

$$\mathbf{Ax} = \mathbf{y}$$

If $\mathbf{A}$ is the identity matrix, then, of course, the vector $\mathbf{x}$ equals the vector $\mathbf{y}$ in the sense that all the components have the same value. Also, if $\mathbf{x} = 0$, then $\mathbf{y} = 0$, but in the future we shall exclude the function (vector) that is identically zero.

Usually the output vector $\mathbf{y}$ will point in a different direction (in the N-dimensional space) from the input vector $\mathbf{x}$. For the typical matrix $\mathbf{A}$ of dimension N, there will be N different vectors $\mathbf{x}$ such that the corresponding $\mathbf{y}$ will have the *same direction* as did the $\mathbf{x}$, although not necessarily the same length. That is, we will have

$$\mathbf{Ax} = \lambda\mathbf{x}$$

for some constant λ. To see the truth of this remark, we can write the preceding equation in the form

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

where $\mathbf{I}$ is the identity matrix. For this equation to have a solution that is not identically zero, it is both necessary and sufficient that the determinant of the system of equations

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

This determinant,

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} - \lambda & a_{23} & \cdots \\ a_{31} & a_{32} & a_{33} - \lambda & \cdots \\ \vdots & \vdots & \vdots & \end{vmatrix} = 0$$

when expanded, is clearly a polynomial in λ of degree N, and, in general,

it will have N distinct zeros $\lambda_1, \lambda_2, \ldots, \lambda_N$, real or complex (to have a multiple zero is a restriction on the matrix A). Thus there are, in general, N distinct λ_k with corresponding vector solutions $\mathbf{x}_k$. (Note that $\mathbf{x}_k$ is a vector and not a component of a vector.) The values λ_k are called the *eigenvalues*, and the $\mathbf{x}_k$ are called the corresponding *eigenvectors*. For a given eigenvalue, the determinant is zero, and the corresponding eigenvector is, of course, determined only to within a multiplicative constant.

Why are these eigenvectors important? There are (almost always) N distinct eigenvectors, and they can be shown to be *linearly independent*. Therefore, they can serve as a basis for representing an arbitrary vector $\mathbf{x}$ of N dimensions. Thus we can represent an arbitrary vector $\mathbf{x}$ as a linear combination of the N eigenvectors $\mathbf{x}_k$:

$$\mathbf{x} = \sum_{k=1}^{N} a_k \mathbf{x}_k$$

If we now multiply this equation on both sides by the matrix A (technically, we apply the operation A to the equation), we find that

$$\mathbf{A}\mathbf{x} = \sum_{k=1}^{N} a_k \mathbf{A}\mathbf{x}_k = \sum_{k=1}^{N} a_k \lambda_k \mathbf{x}_k$$

and we see that each eigenvector is multiplied by its corresponding eigenvalue. In the eigenvector representation, the effect of the multiplication by the matrix A (applying the operation A) is easy to follow. The eigenvectors are independent of each other.

To state this important property in other words, we can say that the eigenfunctions do not "feel" the presence of each other; each minds its own business regardless of how much there may or may not be of the other eigenfunctions.

To illustrate the above consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$$

This leads to the corresponding determinant

$$|\mathbf{A} - \lambda \mathbf{I}| = \begin{vmatrix} 1 - \lambda & 2 \\ 3 & 2 - \lambda \end{vmatrix} = 0$$

Upon expansion of the determinant, we get the equation for the eigenvalues

$$\lambda^2 - 3\lambda + 2 - 6 = 0$$

which has zeros $\lambda = 4, -1$. If we use $\lambda_1 = 4$, we obtain for the matrix equation (where we are using the notation $x_{i,j}$ for the jth component of the ith vector $\mathbf{x}_i$)

$$\begin{pmatrix} -3 & 2 \\ 3 & -2 \end{pmatrix} \begin{pmatrix} x_{1,1} \\ x_{1,2} \end{pmatrix} = 0$$

which leads to the single equation

$$-3x_{1,1} + 2x_{1,2} = 0$$

and the corresponding eigenvector

$$\begin{pmatrix} x_{1,1} \\ \frac{3x_{1,1}}{2} \end{pmatrix} = \frac{x_{1,1}}{2} \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

The value of $x_{1,1}$ is arbitrary, since the rank of the matrix for the eigenvalue is 1. If we use the other eigenvalue $\lambda_2 = -1$, we obtain, correspondingly,

$$\begin{pmatrix} 2 & 2 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} x_{2,1} \\ x_{2,2} \end{pmatrix} = 0$$

with the corresponding eigenvector

$$\begin{pmatrix} x_{2,1} \\ -x_{2,1} \end{pmatrix} = x_{2,1} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

These two eigenvectors,

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

are linearly independent and can represent any arbitrary two-dimensional vector.

In particular the general vector

$$\begin{pmatrix} a \\ b \end{pmatrix} = \frac{a+b}{5} \begin{pmatrix} 2 \\ 3 \end{pmatrix} + \frac{3a-2b}{5} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

hence

$$\mathbf{A} \begin{pmatrix} a \\ b \end{pmatrix} = 4 \frac{a+b}{5} \begin{pmatrix} 2 \\ 3 \end{pmatrix} - \frac{3a-2b}{5} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Exercises

2.3-1 Find the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 \\ 2 & 0 \end{pmatrix}$$

2.3-2 Given the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & -2 \\ 1 & 0 & 0 \end{pmatrix}$$

find all the eigenvalues and the eigenvector corresponding to the eigenvalue 1. *Answer:* $(1, -1, 1)$.

2.4 INVARIANCE UNDER TRANSLATION

In many data-processing problems there is no natural origin, and therefore an arbitrary point is selected as the origin (typically, for a time signal, it is the time when we set $t = 0$, which is arbitrary). From the addition formulas of trigonometry

$$\begin{aligned} \sin(x + y) &= \sin x \cos y + \cos x \sin y \\ \cos(x + y) &= \cos x \cos y - \sin x \sin y \end{aligned} \tag{2.4-1}$$

it is an easy exercise in trigonometry to see that, when $x = x' + h$,

$$A \sin x + B \cos x$$

becomes

$$A' \sin x' + B' \cos x'$$

where we have

$$\begin{aligned} A' &= A \cos h - B \sin h \\ B' &= A \sin h + B \cos h \end{aligned} \tag{2.4-2}$$

Squaring each of these expressions and adding them, we obtain

$$A'^2 + B'^2 = A^2 + B^2 \tag{2.4-3}$$

Thus we see that under the operation of translation the *pair* of functions $\cos x$ and $\sin x$ constitutes the eigenfunctions, for when placed into the operation of translation by the amount h, they again emerge.

The famous Euler identities are

$$\cos x + i \sin x = e^{ix}$$
$$\cos x - i \sin x = e^{-ix} \tag{2.4-4}$$

(where $i = \sqrt{-1}$). Note that, for x real, as is assumed,

$$| e^{ix} | = 1 \tag{2.4-5}$$

To verify the equations (2.4-4) we expand e^{ix} in a power series

$$e^{ix} = 1 + (ix) + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \ldots$$

Rearrange and separate the real and imaginary parts, noting that $i^2 = -1$, $i^4 = 1$, etc.

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots\right)$$

and we recognize the corresponding sine and cosine expansions. Thus we have the top equation of (2.4-4). Replace i by $-i$ and you have the lower equation.

The equations (2.4-4) lead to the corresponding formulas

$$\cos x = \frac{1}{2}(e^{ix} + e^{-ix})$$
$$\sin x = \frac{1}{2i}(e^{ix} - e^{-ix}) \tag{2.4-6}$$

In this notation the two addition formulas (2.4-1) from trigonometry are contained in the single, much simpler formula

$$e^{ix}e^{iy} = e^{i(x+y)}$$

This fact can be seen by using the Euler identities (2.4-4) on both sides and then equating the real and imaginary terms on each side. Thus the complex

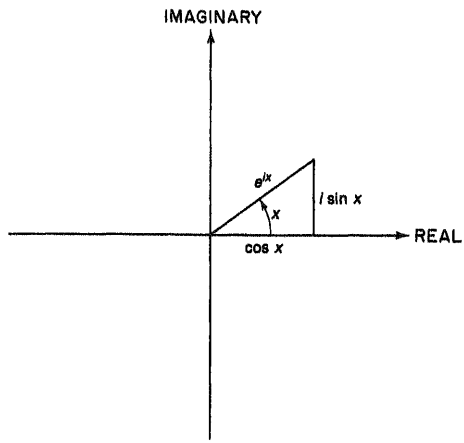


FIGURE 2.4-1

exponentials are the eigenfunctions of translation. *It is far more convenient, therefore, to use the complex exponentials than it is to use the real sines and cosines.* On the other hand, the real functions are familiar, and the complex exponentials have a mysterious aura about them. It is simply necessary to get used to the fact that the complex exponentials are the real functions in a slight disguise. You can think of the cosine and sine as the components of a vector e^{ix}. See Figure 2.4-1.

We have chosen the mathematical convention of $i = \sqrt{-1}$ rather than the engineering convention of j. The choice is arbitrary, but since the book is designed to be read by nonengineers, i is a reasonable choice, and engineers need to be familiar with both notations.

It is natural to ask if the sine and cosine are unique in having this property of invariance under translation. The invariant property that we want is that both functions under a translation of a fixed amount, say h, can be written as a linear combination of sine and cosine or, more generally, that any linear combination of a sine and cosine of a given frequency can be written as a linear combination when an arbitrary translation of size h of the coordinate axis is made. We are further assuming that the functions are odd and even, respectively, and that the trigonometric functions, sine and cosine, are reasonably smooth. Under these assumptions it can be shown that the sines and cosines are unique (except that the corresponding hyperbolic functions are also possible). Notice that in the complex exponential notation the eigenfunction property is much more simply expressed than in the trigonometric notation,

$$u(t + h) = e^{i\omega(t+h)} = e^{i\omega h}e^{i\omega t} = \lambda(\omega)u(t)$$

where $\lambda(\omega)$ is the eigenvalue

$$\lambda(\omega) = e^{i\omega h}$$

and is independent of the variable t.

Thus the $e^{i\omega t}$ are the eigenfunctions of translation.

Corresponding to equation (2.4-1), we have in the complex notation

$$u(t) = ce^{i\omega t}$$

(where c can be a complex number) the invariant (2.4-3)

$$(ce^{i\omega t})(\bar{c}e^{-i\omega t}) = c\bar{c} = |c|^2 \tag{2.4-7}$$

2.5 LINEAR SYSTEMS

The second eigenfunction property that we wish to show is that the complex exponential functions $e^{i\omega t}$ and $e^{-i\omega t}$ are eigenfunctions for linear, time-invariant systems. In abstract notation this means

$$L\{e^{i\omega t}\} = \lambda(\omega)e^{i\omega t}$$

where $L\{\ \}$ is an arbitrary linear time invariant operator and $\lambda(\omega)$ does not depend on t. A linear operator has the property that

$$L\{ag_1(t) + bg_2(t)\} = aL\{g_1(t)\} + bL\{g_2(t)\}$$

For example, integration, differentiation, and interpolation are all linear operators. Clearly, for nonrecursive filters of the form

$$y_n = \sum_{k=-N}^{N} c_k u_{n-k}$$

the substitution

$$u(t) = e^{i\omega t}$$

produces, when we factor out the exponential term depending on n, the output

$$y(n) = e^{i\omega n} \sum_{k=-N}^{N} c_k e^{-i\omega k} = \lambda(\omega)e^{i\omega n}$$

where

$$\sum_{k=-N}^{N} c_k e^{-i\omega k} = \lambda(\omega) \tag{2.5-1}$$

Thus the function $e^{i\omega t}$, which we put into the right side of the equation, can be factored out of the expression and appears multiplied by its eigenvalue $\lambda(\omega)$. The eigenvalue $\lambda(\omega)$ is, of course, a constant as far as t or, equivalently, n is concerned, and is usually called the *transfer function*. Thus the transfer function is just the eigenvalue corresponding to the eigenfunction $e^{i\omega t}$.

For a recursive filter we need to assume that *both* the input and the output are of the form

$$u_n = AIe^{i\omega n} \quad \text{and} \quad y_n = AOe^{i\omega n}$$

where the input coefficient AI and the output coefficient AO may be complex numbers. We get from equation (1.1-5)

$$AOe^{i\omega n} = AI \sum_{0}^{N} c_k e^{i\omega(n-k)} + AO \sum_{1}^{M} d_k e^{i\omega(n-k)}$$

or the ratio of output to input coefficients

$$\frac{AO}{AI} = \frac{\sum_{0}^{N} c_k e^{-i\omega k}}{1 - \sum_{1}^{M} d_k e^{-i\omega k}} \tag{2.5-2}$$

This is a transfer function because, when you multiply the input by this ratio you get the output; you "transfer" from the input to the output.

It is worth noting that the exponential function is also the eigenfunction that is appropriate for the calculus operations of differentiation,

$$\frac{d}{dt} e^{i\omega t} = i\omega e^{i\omega t} \tag{2.5-3}$$

and integration,

$$\int e^{i\omega t}\, dt = \frac{e^{i\omega t}}{i\omega} \tag{2.5-4}$$

The exponential is also the eigenfunction for differencing, since

$$\Delta e^{i\omega t} = e^{i\omega(t+1)} - e^{i\omega t} = e^{i\omega t}[e^{i\omega} - 1]$$

Thus we see, contrary to the impression gained from the usual calculus course, that the powers of x are not the eigenfunctions of calculus. Instead, the exponentials, real or complex, are the natural, the characteristic, the eigenfunctions of the calculus.

The expression $e^{i\omega}$ occurs frequently and it is often easier to make a notational change and write

$$e^{i\omega} = z$$

Hence (2.5-1) has the form

$$\sum_{k=-N}^{N} c_k z^{-k} = \lambda(\omega)$$

and (2.5-2) has the form

$$\frac{AO}{AI} = \frac{\sum_{0}^{N} c_k z^{-k}}{1 - \sum_{1}^{M} d_k z^{-k}}$$

Exercises

2.5-1 Find the eigenvalue corresponding to the kth derivative.

2.5-2 Find the eigenvalue corresponding to the kth difference operator Δ^k.

2.5-3 Discuss the lack of an additive constant in (2.5-4).

2.6 THE EIGENFUNCTIONS OF EQUALLY SPACED SAMPLING

The purpose of this section is to show that the eigenfunctions of the process of equally spaced sampling of a function are the common sines and cosines of trigonometry (or equally the complex exponentials e^{ix} and e^{-ix}). The sense in which we mean that they are eigenfunctions is that when we (1) take a sinusoid of some frequency (think of it as a high frequency), then

(2) do the process of sampling the given frequency function at equally spaced points, and (3) finally ask "What equivalent sinusoid of low frequency do we have?" we find that it is equivalent to a *single* sinusoid function. Stated simply, aliasing takes any particular frequency and, in the sense of having the same values at the sample points, transforms it into a single low-frequency function.

Let us contrast this result with what happens when the classical polynomial method of approximation is used. In polynomial approximation when we use the sample points x_i ($i = 1, \ldots, N$), we are led directly to consider the *sample polynomial* defined by

$$\pi(x) = [x - x_1][x - x_2] \cdots [x - x_N]$$

This function plays a central role in the theory, because it is the function that vanishes at all the sample points x_i and thus is the function that we cannot "see." Now, given any power of x, say x^m, we divide this power by $\pi(x)$ in order to get a quotient $Q(x)$ and a remainder $R(x)$:

$$x^m = \pi(x)Q(x) + R(x)$$

where $R(x)$ is of degree less than N. A simple generalization of the standard remainder theorem shows that *at the sample points* x_i the two functions x^m and $R(x)$ have exactly the same values. Thus the original single power of x is aliased into a polynomial $R(x)$, which is, of course, a linear combination of $1, x, x^2, \ldots, x^{(N-1)}$, and not a single power. In this sense, the powers of x are not eigenfunctions for sampling at any spacing. Aliasing for polynomials is a messy business.

Let us restate this result. If we regard the process as (1) starting with a basis function (a power of x, say x^m), (2) sampling at N points, and finally, (3) constructing from the samples a new function of minimal degree in x, then we see that, in general, a single power of x does not go into a power of x. On the other hand, for sinusoids, the process of equally spaced sampling followed by the reconstruction of a function of minimal frequency does result in a single sinusoid. Consequently, in this sense, the sinusoids are the eigenfunctions of equally spaced sampling, and the process reveals once more the central role that aliasing plays in the equally spaced sampling process.

2.7 SUMMARY

In this chapter we have discussed the phenomenon of aliasing, which is due solely to the equally spaced sampling of the original signal. We have also given three reasons why the trigonometric functions, sine and cosine,

are to be used in many filter problems as the basis of representing signals. The reasons are that they are the eigenfunctions for (1) invariance under translation by an arbitrary amount, (2) linear systems, and (3) equally spaced sample systems.

In the complex form $e^{i\omega t}$ and $e^{-i\omega t}$, the trigonometric functions are more easily handled in many problems. Unfortunately, most students believe that we are modeling a real world, and they believe that in the final analysis they have to deal with real signals. The complete equivalence of the two forms, real and complex, does not convince them that the two approaches are exactly equivalent.

At any frequency we have both a sine and a cosine as the basis for representation of any function; in the complex notation we have the positive and negative frequencies to use, and thus the same amount of linear independence. Ultimately, the convenience of the complex notation must be mastered, because it also leads more readily to the deeper insights of what is going on with all signal processing.

In recognition of the reality of the prejudice, we will for a time continue to give both the real and complex forms; but finally we will have to settle on the complex notation as our main tool. When you have to deal with a real function, then the coefficients in the complex form are conjugates of each other, which makes the two terms conjugates of each other, and their sum is thus real.

We have also introduced the z- transform

$$z = e^{i\omega}$$

which at present is a mere notational convenience and adds nothing to the theory.