Linear Classification of Neural Manifolds with Correlated Variability

Albert J. Wakhloo,[1,2] Tamara J. Sussman,[2,3] and SueYeon Chung[1,4]

[1] *Center for Computational Neuroscience, Flatiron Institute*
[2] *Department of Child and Adolescent Psychiatry, New York State Psychiatric Institute*
[3] *Columbia University Irving Medical College*
[4] *Center for Neural Science, New York University*

Understanding how the statistical and geometric properties of neural activity relate to performance is a key problem in theoretical neuroscience and deep learning. Here, we calculate how correlations between object representations affect the capacity, a measure of linear separability. We show that for spherical object manifolds, introducing correlations between centroids effectively pushes the spheres closer together, while introducing correlations between the axes effectively shrinks their radii, revealing a duality between correlations and geometry with respect to the problem of classification. We then apply our results to accurately estimate the capacity of deep network data.

Introduction: Neural networks can learn rich representations of the world. This capacity for representation learning is thought to underlie deep learning's unprecedented success across a wide variety of tasks. However, it is unclear how the geometric and statistical properties of neural network representations shape network performance on common tasks. Recent work addresses this gap by studying the interaction between artificial neural network representations and performance on classification and memorization tasks [1–10], with complementary work in neuroscience studying the interaction between the structure of biological neural network representations and animal behavior [11–14]. Specifically, in [15–17], the authors introduce the manifold shattering capacity, a measure capturing how easy it is to separate random binary partitions of a set of manifolds with a hyperplane, and express it in terms of the underlying manifold geometry. In this way, network performance on a classification task, as measured by the capacity, can be understood through the geometric structure of the network representations.

Previous works on the manifold capacity have either ignored or coarsely approximated the effects of neural correlations. The best approximation to these effects was reported in [16], where the authors "project out" low-rank correlation structures in manifold centroids. However, the authors find that this approach breaks down when applied to certain artificial network data. Moreover, this approach does not offer analytical insight into the role of different types of correlations in object classification.

Object representations in artificial and biological neural networks exhibit intricate correlation structures, which reflect important properties of the underlying representations [22–25]. Moreover, as the deep learning community shifts to a self-supervised learning paradigm, many popular loss functions directly enforce particular correlation structures between the latent representations of (possibly augmented) batches of data points [26–29]. These considerations call for a theoretical characterization of the relationship between network performance, representational geometry, and the correlation structure of network representations.

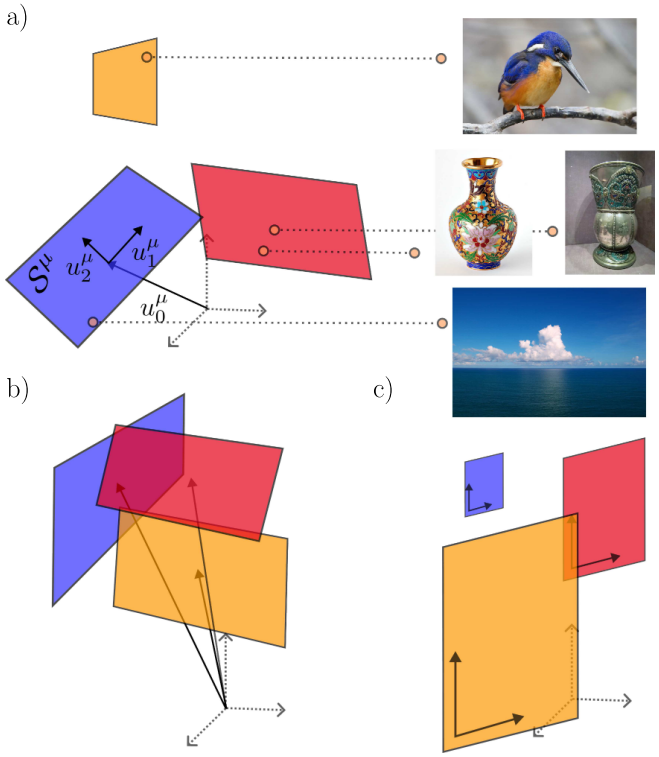

FIG. 1. (a) Neural manifolds arising from different instances of $P = 3$ object classes (bird, vase, and cloud [18–21]), with $N = 3$ neurons. We parametrize the manifolds in terms of a centroid u_0^μ, axes $u_{i>0}^\mu$, and shape vectors $\mathcal{S}^\mu$, determining which linear combinations of the axes lie within the manifold. (b) Neural manifolds with correlations in their centroids. (c) Neural manifolds with fully correlated axes. In all three images, different colors correspond to different object class manifolds.

In this Letter, we calculate the effects of correlation structures on the capacity. Our formula for the capacity of correlated manifolds generalizes the results in [15] by stretching the Euclidean norm appearing in previous results in the directions of the eigenvectors of the covariance tensor. We analyze this formula in a simple setting,

showing how geometry and correlations interact to determine the capacity, and we go on to apply this formula to accurately estimate the capacity of deep network data.

Problem statement: Consider a set of P manifolds, M^μ, residing in $\mathbb{R}^N$. These manifolds correspond to distinct sets of neuronal activation vectors when presented with different types of stimuli—for example, the set of neural activations for a set of P classes across all possible class instances in a given layer of an image recognition network [Fig. 1(a)]. In what follows, we assume that each manifold resides in an affine subspace of maximal dimension $K < N$. That is, for any $x \in M^\mu$, we have that $x = u_0^\mu + \sum_{i=1}^K s_i^\mu u_i^\mu$, where u_0^μ is a manifold center, u_i^μ for $1 \le i \le K$ is a set of manifold axes, and $s \in \mathcal{S}^\mu$ are the coordinates of x with respect to the manifold axes. We use $\mathcal{S}^\mu \subset \mathbb{R}^K$ to denote the set of all possible coordinates in this basis.

We take the manifold center u_0^μ to be the average activation of the network layer when presented with a data point from class μ. The spread of the manifold along the axes therefore corresponds to the network variability as we sample different stimuli from class μ. Intuitively, manifolds with large centroid norms far away from one another with small spreads along their axes will be easier to classify than large manifolds tightly packed together.

We now turn to the problem of determining the maximal number of manifolds per dimension, $\alpha \equiv P/N$, which are, given some random binary labelings $y^\mu \in \{-1, 1\}$ and some underlying distribution on the u_i^μ, linearly separable with high probability at a fixed margin κ. In what follows, we will be specifically interested in the thermodynamic limit, $N, P \to \infty$ with $P/N = O(1)$. In other words, we find the greatest α such that there exists a hyperplane with normal $w \in \mathbb{R}^N$, $||w||_2^2 = N$ satisfying $\min_{x \in M^\mu} y^\mu \langle w, x \rangle \ge \kappa$ for each manifold M^μ with probability 1 in this limit. We define the manifold capacity to be this maximal value of α, so that larger capacities imply a more favorable representational geometry for the purpose of classification.

Following [2, 15, 30–34], we study this problem by calculating the average log-volume of the space of solutions in the thermodynamic limit:

$$\overline{\log \mathrm{Vol}} = \log \int_{\mathbb{S}(\sqrt{N})} d^N w \prod_\mu \Theta\Big(\min_{x \in M^\mu} y^\mu \langle w, x \rangle - \kappa \Big) , \tag{1}$$

where $\mathbb{S}(\sqrt{N})$ is the sphere of radius $\sqrt{N}$, $\Theta(\cdot)$ is the Heaviside step function, and the average is taken with respect to the quenched disorder in the labels y^μ and the axes and centroids u_i^μ. Viewing the volume as a partition function, we can see that $-N^{-1} \log \mathrm{Vol}$ corresponds to a free energy density, which we assume is self-averaging [35]. Given a fixed set of manifold shapes $\mathcal{S}^\mu$, and choosing the axes and centroids to be independent from one another with $u_i^\mu \sim \mathcal{N}(0, N^{-1} I^{(N)})$, the capacity for such randomly oriented manifolds, α_M, is given by [15]

$$\frac{1}{\alpha_M(\kappa)} = \frac{1}{P} \int D_I T \min_{V \in \mathcal{A}} \sum_{i,\mu} (V_i^\mu - T_i^\mu)^2 , \tag{2}$$

where $D_I T = \prod_{\mu,i} dT_i^\mu \exp[-\frac{1}{2}(T_i^\mu)^2]/\sqrt{2\pi}$ is an isotropic Gaussian measure and $\mathcal{A}$ is a convex set of matrices which depends on the geometry of the manifolds, as reflected by their shapes, $\mathcal{S}^\mu$:

$$\mathcal{A} \equiv \Big\{ V \in \mathbb{R}^{P \times (K+1)} : V_0^\mu + \min_{s^\mu \in \mathcal{S}^\mu} \sum_{i=1}^K V_i^\mu s_i^\mu \ge \kappa \Big\}. \tag{3}$$

Note the similarity to the constraint in the Θ function in Eq. (1) . Indeed, the variable V_i^μ corresponds to the inner product of the solution vector w with the ith axis (or centroid) of the μth manifold, multiplied by the label: $V_i^\mu \equiv y^\mu \langle w, u_i^\mu \rangle$. These are the so-called signed fields of the solution vector on the u_i^μ [15]. In this way, the capacity can be understood as a function of the geometry of the manifolds as reflected in the set $\mathcal{S}^\mu$. In the special case that the manifolds are simply randomly oriented points, the capacity is given by [30]

$$\frac{1}{\alpha_{point}(\kappa)} = \int_{-\infty}^{\kappa} \frac{d\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} (\xi - \kappa)^2 . \tag{4}$$

From this formula, we can see that the shape sets $\mathcal{S}^\mu$ cause a lower capacity when compared to that of points.

Replica theory for correlated manifolds: Here, we consider the situation where manifold axes and centroids are correlated with one another. Intuitively, this corresponds to the fact that different classes in a dataset may be more or less similar to one another in the neural representation space. We enforce correlated axes and centroids by assuming that $\overline{\langle u_i^\mu, u_j^\nu \rangle} = C_{\nu,j}^{\mu,i}$ for some positive definite covariance tensor $C_{\nu,j}^{\mu,i}$. This is done by placing a Gaussian distribution on the centroids and axes: $p(u) \propto \exp\big[-\frac{N}{2} \sum_{\mu,\nu,i,j,l} (C^{-1})_{\nu,j}^{\mu,i} u_{i,l}^\mu u_{j,l}^\nu \big]$.

We calculate the capacity for correlated manifolds using the replica method [35, 36]; the details can be found in the Supplementary Material (SM) [37]. We find that the capacity at a margin κ, denoted by $\alpha_{cor}(\kappa)$, is

$$\frac{1}{\alpha_{cor}(\kappa)} = \frac{1}{P} \overline{\int D_{y,C} T \min_{V \in \mathcal{A}} ||V - T||_{y,C}^2} , \tag{5}$$

where $D_{y,C} T$ is the zero-mean Gaussian measure with covariance tensor $y^\mu y^\nu C_{\nu,j}^{\mu,i}$, and the overline denotes the remaining average with respect to the labels y^μ. Note too that we have defined the Mahalanobis norm: $||X||_{y,C}^2 \equiv \sum_{\mu,\nu,i,j} y^\mu y^\nu \big(C^{-1}\big)_{\nu,j}^{\mu,i} X_i^\mu X_j^\nu$, which effectively stretches

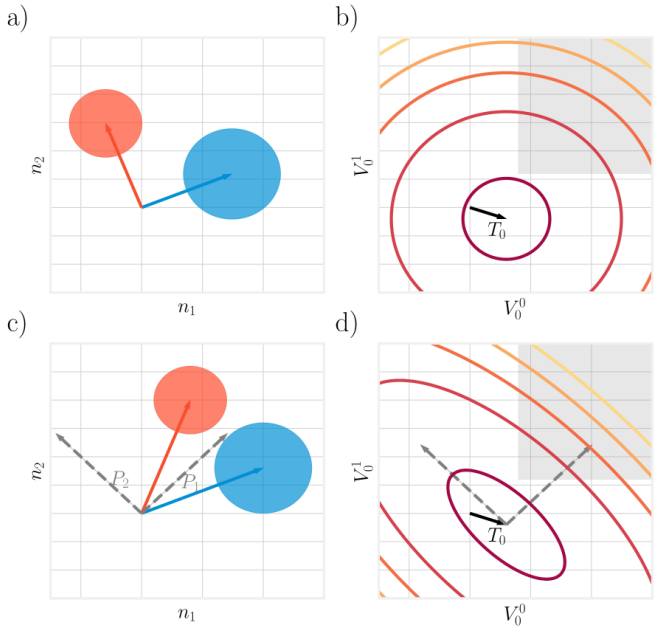

FIG. 2. The effect of correlations on the optimization landscape for V_0. *First column:* Two manifolds with (a) uncorrelated and (c) correlated centroids arising from the activations of two neurons, n_1 and n_2. *Second column:* Level curves for $||V-T||^2_{y,C}$, given fixed y and $V_{i>0}$ for the (b) uncorrelated and (d) correlated manifolds. Shaded regions correspond to areas where the constraint is satisfied—i.e., sections of the set $\mathcal{A}$ in Eq. (3). Clearly, correlations warp the optimization landscape along the eigenvectors P_1, P_2 of the centroid covariance matrix with off-diagonal sign flips $y^\mu y^\nu C^{\mu,0}_{\nu,0}$.

the Frobenius norm along the eigenvectors of the tensor $y^\mu y^\nu C^{\mu,i}_{\nu,j}$ (Fig. 2).

Comparison with other capacity estimators: It is worth pausing and comparing Eq. (5) to the solution for uncorrelated manifolds in Eq. (2) reported in [15, 16]. From Eqs. (2) and (5), we can see that axes and centroid correlations distort the norm in the minimization from the Euclidean norm to a random Mahalanobis norm which depends on the covariance tensor C and the random labels y^μ (Fig. 2). As such, we expect that the quality of the α_M estimator from Eq. (2) degrades as the manifold axes and centroids become more correlated with one another. We find that this is the case for both α_M and the low-rank approximation method reported in [16] when applied to Gaussian point cloud manifolds (Fig. 3). Therefore, the correlated capacity estimator, α_{cor}, whose numerical implementation we describe in the SM [37], should be used whenever working with manifolds with strong correlations (see [39]).

The special case of spheres: We now look for an answer to the problem we were originally interested in: What are the effects of manifold correlations on the capacity? We answer this question by analytically solving Eq. (5) in a simple setting: K-dimensional spheres with homogeneous axis and centroid correlations. More precisely, we assume that the manifold shape sets $\mathcal{S}^\mu$ are spheres of radius 1, and the covariance tensor C is defined by

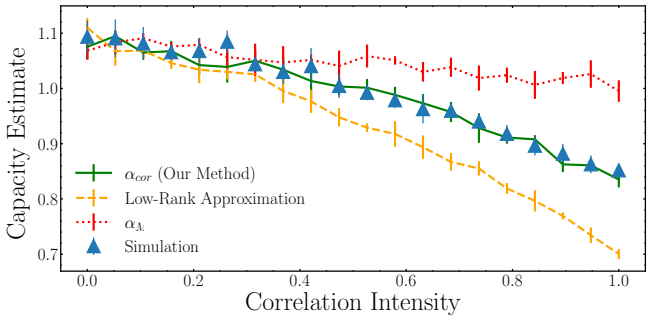

FIG. 3. Comparison of three different capacity estimators, including the low-rank approximation of [16], to the numerically estimated ground truth simulation capacity (blue triangles) described in [40]. The correlation intensity denotes the magnitude of the off-diagonal correlations—see SM for more details [37].

$$C^{\mu,i}_{\nu,j} \equiv \begin{cases} \delta_{i,j}[(1-\lambda)\delta_{\mu,\nu}+\lambda] & \text{for } i,j>0 \\ (1-\psi)\delta_{\mu,\nu}+\psi & \text{for } i,j=0 \\ 0 & \text{for } i>0, j=0\,, \end{cases} \tag{6}$$

where $0 \le \psi, \lambda < 1$. The average centroid norms and sphere radii are then respectively controlled by the scalars r_0 and r, so that for all μ and $x \in M^\mu$, we have that $x = r_0 u^\mu_0 + r\sum_{i=1}^K s_i u^\mu_i$, with $\sum_i (s_i)^2 \le 1$. The variables λ, ψ respectively determine the degree of correlation between the axes and centroids: As $\lambda, \psi \to 1$, the axes and centroids will be fully correlated with one another, while $\lambda, \psi \to 0$ implies randomly oriented axes and centroids [Fig. 4(a)].

Even under these simplifying assumptions, the minimization in Eq. (5) is not directly solvable. As such, we reframe the problem in terms of a statistical mechanical system with quenched disorder and study the limit $P \to \infty$. To do this, note that the constraint on the fields can be rewritten as $r_0 V^\mu_0 - r\sqrt{\sum_{i>0}(V^\mu_i)^2} \ge \kappa$, as can be seen by applying the Karush-Kuhn-Tucker (KKT) conditions [41] to the Lagrangian $\mathcal{L}(S,\eta) = r\sum_{i>0} V^\mu_i S_i + \eta(||S||^2 - 1)$ (see SM [37]). The capacity can then be derived by studying the following Gibbs measure:

$$\frac{1}{Z}\exp\left[-\frac{\beta}{2}\sum_{i,j,\mu,\nu} y^\mu y^\nu (C^{-1})^{\mu,i}_{\nu,j}(V^\mu_i - T^\mu_i)(V^\nu_j - T^\nu_j)\right] \times \prod_\mu \Theta\left(r_0 V^\mu_0 - r\sqrt{\sum_{i>0}(V^\mu_i)^2} - \kappa\right) dV^\mu\,, \tag{7}$$

where Z is the partition function [42]. We can see that $1/\alpha_{cor}(\kappa)$ is then given by the average energy in the zero-temperature limit: $[\alpha_{cor}(\kappa)]^{-1} = -\frac{2}{P}\lim_{\beta\to\infty}\frac{\partial}{\partial\beta}\overline{\log Z}$,

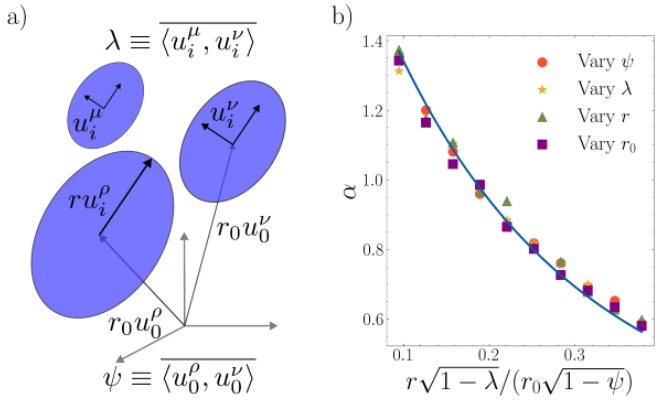

FIG. 4. The capacity for correlated spheres. (a) Visual demonstration of spherical manifolds with low-rank axis and centroid correlations. (b) The zero-margin capacity as a function of only the input ratio $r\sqrt{1-\lambda}/(r_0\sqrt{1-\psi})$. Points represent averages over five random sphere samplings, and the solid line represents the theoretical prediction. For each experiment, we fix three of the four parameters and vary the remaining one to obtain a fixed value of the ratio.

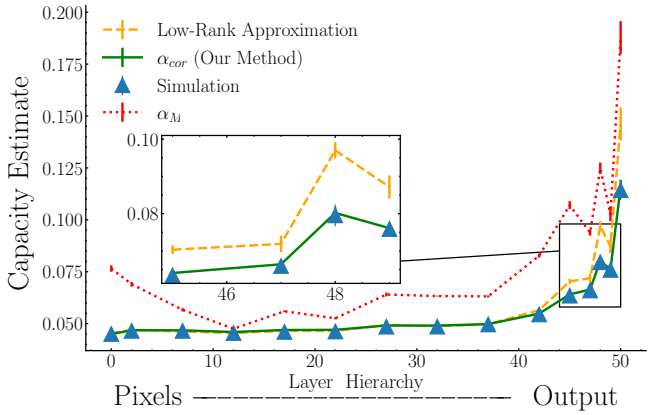

FIG. 5. Comparison of the low-rank approximation (yellow dashed line) [16], α_M (red dotted line) [15], and our α_{cor} calculation (green solid line) to the ground truth simulation capacity (blue triangles) [17] on data manifolds arising from the ResNet50 artificial neural network architecture trained using SimCLR on the ImageNet dataset [26, 43].

with the overline denoting the average with respect to the T and the labels y^μ. We calculate the resulting free energy density using the replica method—see SM for details [37].

Under these assumptions, the capacity is given by

$$\frac{1}{\alpha_{cor}(\kappa)} = K\left(\sqrt{q}-1\right)^2 + \int_{-\infty}^{\hat{\kappa}(q)} \frac{d\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \left(\xi - \hat{\kappa}(q)\right)^2, \tag{8}$$

where q is the scaled squared norm of the signed fields of an arbitrary sphere, $q \equiv \overline{\sum_{i>0}(V_i^\mu)^2}/(K(1-\lambda))$, and $\hat{\kappa}(q)$ is an effective margin. The values of the q and $\hat{\kappa}(q)$ are then fixed by the self-consistent equations

$$\sqrt{q} = 1 + \frac{r\sqrt{1-\lambda}}{r_0\sqrt{K(1-\psi)}} \int_{-\infty}^{\hat{\kappa}(q)} \frac{d\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \left(\xi - \hat{\kappa}(q)\right),$$
$$\hat{\kappa}(q) = \frac{r\sqrt{K(1-\lambda)q} + \kappa}{r_0\sqrt{1-\psi}}. \tag{9}$$

With our definition of q and $\hat{\kappa}(q)$ in hand, we can see that the capacity for correlated spheres is the same as the capacity of random points given in Eq. (4) with an effective margin of $\hat{\kappa}(q)$, plus an extra bias term which corresponds to additional contributions to the capacity from the correlations and spread of the spheres.

The above solution gives a direct view into the effects of correlations on manifold separability. From Eqs. (8) and (9), we can see that when $\kappa = 0$, both q and the effective margin are fully determined by the ratio $r\sqrt{(1-\lambda)}/(r_0\sqrt{1-\psi})$ (Fig. 4). Even when $\kappa \neq 0$, the sphere radii and centroid scalings, r, r_0, and the respective correlations, λ, ψ, only affect the capacity through the products: $r\sqrt{1-\lambda}$, $r_0\sqrt{1-\psi}$. This implies that increasing the axis or centroid correlations affects the capacity in the same way as shrinking the spheres or centroid norms does. That is, axis correlations effectively shrink the sphere radii, while centroid correlations effectively push the manifolds closer to the origin.

These effects are most dramatic when we consider the limits of fully correlated manifolds. In the fully correlated centroids limit, $\psi \to 1$, we can see that the capacity falls to 0. Conversely, in the fully correlated axes limit, $\lambda \to 1$, we can see that $\sqrt{q} \to 1$, so that the capacity grows to the capacity for random points with margin $\kappa/(r_0\sqrt{1-\psi})$ [30]. This shows that high-dimensional, fully correlated spheres are as easy to separate as randomly oriented points—see [4, 34] for related results.

Application to deep network manifolds: Having studied our theoretical predictions in two simple settings, we now consider the performance of our capacity estimator, α_{cor}, when applied to neural manifolds from a pretrained SimCLR ResNet50 network on the ImageNet dataset [26, 43, 44]. We can see from Fig. 5 that the low rank approximation [16] significantly overestimates the capacity in later layers of the network. Note that while we can numerically estimate the ground truth simulation capacity here because we use few data points (see SM; [37]), this is computationally infeasible for larger data manifolds [16, 40]. Thus, our α_{cor} estimator can be used to estimate the capacity where other methods fail.

Discussion: In this Letter, we considered the problem of linearly separating a set of high-dimensional manifolds whose centroids and axes are correlated with one another. We first derived an expression for the capacity of general manifolds with arbitrary covariance tensors. After showing that the resulting expression outperforms previ-

ous capacity estimators when presented with correlated manifolds, we turned to the problem of interpreting the resulting expression for the capacity. To this end, we considered the problem of linearly separating spheres with homogeneous correlations along the centroids and axes. The resulting expression for the capacity closely tracks the capacity for points with an effective margin determined by the geometry and correlations of the spheres. Remarkably, we found that centroid and axis correlations play the same roles as the distance of the spheres from the origin and the sphere radii, respectively. These findings reveal a duality between representational geometry and correlations with respect to the problem of classification.

Our work suggests two main subsequent lines of research. First, given the rising popularity and sophistication of geometric analysis methods in neuroscience [11–13, 15, 16], together with the extensive literature examining the phenomenology and role of different types of neural correlations [22, 23], we hope to apply the results from this study to further connect these two lines of inquiry. One particularly interesting approach in this direction would be to apply our results to study the relationship between hierarchical correlation structures, geometry, and the organization of abstract knowledge, especially in the context of multilabel classification [12, 45, 46]. Another interesting approach would be to use Eq. (5) to derive a set of metrics quantifying the effects of different types of neural correlations on the capacity for arbitrary data manifolds, complementing pre-existing measures describing the impact of geometry on the capacity [15, 16].

Second, our results regarding spheres with correlated axes suggest that self-supervised objectives which produce positive correlations between manifold axes could yield latent representations with favorable classification properties. If we further define manifold axes using the translation between an original image and its augmentation, such an objective could also produce representations which are disentangled with respect to, for example, color distortion and rotation [47, 48]. We hope to pursue this line of research in subsequent work.

Acknowledgments: The authors thank Abdulkadir Canatar and Chi-Ning Chou for their comments on an earlier version of this manuscript.

[1] P. Rotondo, M. C. Lagomarsino, and M. Gherardi, Counting the learnable functions of geometrically structured data, Physical Review Research **2**, 023169 (2020).

[2] A. Battista and R. Monasson, Capacity-Resolution Trade-Off in the Optimal Learning of Multiple Low-Dimensional Manifolds by Attractor Neural Networks, Physical Review Letters **124**, 048302 (2020).

[3] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the Influence of Data Structure on Learning in Neural Networks: The Hidden Manifold Model, Physical Review X **10**, 041044 (2020).

[4] M. Farrell, B. Bordelon, S. Trivedi, and C. Pehlevan, Capacity of group-invariant linear readouts from equivariant representations: How many objects can be linearly classified under all possible views?, in *International Conference on Learning Representations* (2022).

[5] T. Biswas and J. E. Fitzgerald, Geometric framework to predict structure from function in neural networks, Physical Review Research **4**, 023255 (2022).

[6] L. Susman, F. Mastrogiuseppe, N. Brenner, and O. Barak, Quality of internal representation shapes learning performance in feedback neural networks, Physical Review Research **3**, 013176 (2021).

[7] A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, in *Advances in neural information processing systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).

[8] D. Dahmen, M. Gilson, and M. Helias, Capacity of the covariance perceptron, Journal of Physics A: Mathematical and Theoretical **53**, 354002 (2020).

[9] J. Steinberg and H. Sompolinsky, Associative memory of structured knowledge, Scientific Reports **12**, 21808 (2022).

[10] U. Cohen and H. Sompolinsky, Soft-margin classification of object manifolds, Physical Review E **106**, 024126 (2022), arXiv:2203.07040 [cond-mat, q-bio, stat].

[11] R. Chaudhuri, B. Gerçek, B. Pandey, A. Peyrache, and I. Fiete, The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep, Nature Neuroscience **22**, 1512 (2019).

[12] S. Bernardi, M. K. Benna, M. Rigotti, J. Munuera, S. Fusi, and C. D. Salzman, The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex, Cell **183**, 954 (2020), publisher: Elsevier.

[13] S. Chung and L. Abbott, Neural population geometry: An approach for understanding biological and artificial neural networks, Current opinion in neurobiology **70**, 137 (2021).

[14] B. Sorscher, S. Ganguli, and H. Sompolinsky, Neural representational geometry underlies few-shot concept learning, Proceedings of the National Academy of Sciences **119**, e2200800119 (2022).

[15] S. Chung, D. D. Lee, and H. Sompolinsky, Classification and Geometry of General Perceptual Manifolds, Physical Review X **8**, 031003 (2018).

[16] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, Separability and geometry of object manifolds in deep neural networks, Nature Communications **11**, 746 (2020), number: 1 Publisher: Nature Publishing Group.

[17] S. Chung, D. D. Lee, and H. Sompolinsky, Linear Readout of Object Manifolds, Physical Review E **93**, 060301 (2016), arXiv:1512.01834 [cond-mat, q-bio, stat].

[18] J. J. Harrison, Azure Kingfisher (2011), `https://upload.wikimedia.org/wikipedia/commons/7/72/Alcedo_azurea_-_Julatten.jpg` This work is licensed under the Creative Commons 3.0 Unported License. To view a copy of this license, visit `https://creativecommons.org/licenses/by/3.0/legalcode`.

[19] S. Korneev, Faberge Vase (2021), `https://upload.wikimedia.org/wikipedia/commons/9/9e/Faberge_vase_State_Museum_of_Sport_1928.jpg` This work is

licensed under the Creative Commons 4.0 ShareAlike License International. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/legalcode.

[20] A. Karwath, Hand-made Chinese Vase (2005), https://upload.wikimedia.org/wikipedia/commons/b/b8/Chinese_vase.jpg This work is licensed under the Creative Commons 2.5 Generic ShareAlike License. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/2.5/legalcode.

[21] T. Fioreze, Clouds over the Atlantic Ocean (2008), https://upload.wikimedia.org/wikipedia/commons/e/e0/Clouds_over_the_Atlantic_Ocean.jpg This work is licensed under the Creative Commons ShareAlike 3.0 Unported License. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/3.0/legalcode.

[22] S. Panzeri, M. Moroni, H. Safaai, and C. D. Harvey, The structures and functions of correlations in neural population codes, Nature Reviews Neuroscience **23**, 551 (2022).

[23] J. Zylberberg, A. Pouget, P. E. Latham, and E. Shea-Brown, Robust information propagation through noisy neural circuits, PLOS Computational Biology **13**, e1005497 (2017).

[24] A. Morcos, M. Raghu, and S. Bengio, Insights on representational similarity in neural networks with canonical correlation, Advances in Neural Information Processing Systems **31** (2018).

[25] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, Similarity of neural network representations revisited, in *International conference on machine learning* (2019) pp. 3519–3529, tex.organization: PMLR.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *International conference on machine learning* (2020) pp. 1597–1607, tex.organization: PMLR.

[27] A. Bardes, J. Ponce, and Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, in *International conference on learning representations* (2022).

[28] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in *International conference on machine learning* (2021) pp. 12310–12320, tex.organization: PMLR.

[29] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, Momentum Contrast for Unsupervised Visual Representation Learning, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, Seattle, WA, USA, 2020) pp. 9726–9735.

[30] E. Gardner, The space of interactions in neural network models, Journal of Physics A: Mathematical and General **21**, 257 (1988).

[31] R. Rubin, R. Monasson, and H. Sompolinsky, Theory of spike timing-based neural classifiers, Physical review letters **105**, 218102 (2010).

[32] F. Schönsberg, Y. Roudi, and A. Treves, Efficiency of Local Learning Rules in Threshold-Linear Associative Networks, Physical Review Letters **126**, 018301 (2021).

[33] R. Monasson, Properties of neural networks storing spatially correlated patterns, Journal of Physics A: Mathematical and General **25**, 3701 (1992).

[34] B. Lopez, M. Schroder, and M. Opper, Storage of correlated patterns in a perceptron, Journal of Physics A: Mathematical and General **28**, L447 (1995).

[35] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (WORLD SCIENTIFIC, 1986) _eprint: https://www.worldscientific.com/doi/pdf/10.1142/0271.

[36] M. Mezard and A. Montanari, *Information, physics, and computation* (Oxford University Press, 2009).

[37] See supplemental material below for details of the replica calculations and experimental details, which includes Ref. [38].

[38] W. W. Hager, Updating the inverse of a matrix, SIAM review **31**, 221 (1989).

[39] A. Wakhloo, T. Sussman, and S. Chung, Capacity for correlated manifolds code, https://zenodo.org/record/7844169#.ZD9Gwy-B22s (2023), 10.5281/zenodo.7844169.

[40] S. Chung, U. Cohen, H. Sompolinsky, and D. D. Lee, Learning data manifolds with a cutting plane method, Neural Computation **30**, 2593 (2018).

[41] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization* (Cambridge university press, 2004).

[42] E. Gardner and B. Derrida, Optimal storage properties of neural network models, Journal of Physics A: Mathematical and general **21**, 271 (1988).

[43] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.

[44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ImageNet large scale visual recognition challenge, International Journal of Computer Vision (IJCV) **115**, 211 (2015).

[45] A. M. Saxe, J. L. McClelland, and S. Ganguli, A mathematical theory of semantic development in deep neural networks, Proceedings of the National Academy of Sciences **116**, 11537 (2019).

[46] W. J. Johnston and S. Fusi, Abstract representations emerge naturally in neural networks trained to perform multiple tasks, Nature Communications **14**, 1040 (2023).

[47] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in *International conference on learning representations* (2017).

[48] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, Towards a Definition of Disentangled Representations, arXiv:1812.02230 [cs, stat] (2018), arXiv: 1812.02230.

Supplementary Material: Linear Classification of Neural Manifolds with Correlated Variability

Albert J. Wakhloo, Tamara J. Sussman, SueYeon Chung

1 Capacity for General Manifolds with Arbitrary Correlations

Claim 1. *Consider a set of manifolds $M^\mu \subset \mathbb{R}^N$ with $\mu = 1, \dots, P$ and with corresponding shape sets $\mathcal{S}^\mu \subset \mathbb{R}^K$. Suppose the axes and centroids $u_i^\mu \in \mathbb{R}^N$ are distributed according to: $p(u) \propto \exp\left[-\frac{N}{2}\sum_{\mu,\nu,i,j,l}(C^{-1})^{\mu,i}_{\nu,j}u^\mu_{i,l}u^\nu_{j,l}\right]$, and assign random binary labels $y^\mu \in \{-1, 1\}$ with equal probability to each manifold. We define the capacity as the maximum number of manifolds per input dimension, $\alpha \equiv P/N$, which admits a solution $w \in \mathbb{S}(\sqrt{N})$ to the separation problem, $\min_{\mu\in\mathbb{N}_1^P}\min_{x\in M^\mu} y^\mu\langle w, x\rangle \geq \kappa$ with probability 1 for $\kappa \geq 0$ in the thermodynamic limit, $N, P \to \infty$, $P/N = O(1)$. Under our assumptions on the manifolds M^μ, the capacity converges to:*

$$\frac{1}{\alpha_{cor}(\kappa)} = \frac{1}{P}\mathbb{E}_y \int D_{y,C}T \min_{V\in\mathcal{A}} ||V - T||^2_{y,C}, \tag{1}$$

where the average is with respect to the i.i.d. labels taking values ± 1 with equal probability, the constraint set is:

$$\mathcal{A} \equiv \left\{ V \in \mathbb{R}^{P\times(K+1)} : \forall \mu \in \mathbb{N}_1^P,\ V_0^\mu + \min_{s\in\mathcal{S}^\mu}\sum_{i>0} V_i^\mu s_i \geq \kappa \right\}, \tag{2}$$

and the Mahalanobis norm $||X||_{y,C}$ is defined by: $||X||^2_{y,C} \equiv \sum_{i,j,\nu,\mu} X_i^\mu X_j^\nu y^\mu y^\nu \left(C^{-1}\right)^{\mu,i}_{\nu,j}$. As in the main text, we also define the Gaussian measure $D_{y,C}T$ as:

$$D_{y,C}T = (2\pi)^{-P(K+1)/2}|G|^{-1/2}\exp\left\{-\frac{1}{2}\sum_{\mu,\nu,i,j} T_i^\mu T_j^\nu y^\mu y^\nu \left(C^{-1}\right)^{\mu,i}_{\nu,j}\right\}\left[\prod_{\mu=1}^P\prod_{i=0}^K dT_i^\mu\right], \tag{3}$$

where $|G|$ is the determinant of the tensor $y^\mu y^\nu C^{\mu,i}_{\nu,j}$, unrolled into a matrix of dimensions $P(K+1)\times P(K+1)$.

Derivation: We calculate the log volume of the space of solutions [1]:

$$\mathbb{E}_{y,u}\log Z \equiv \mathbb{E}_{y,u}\log\int d^N w\delta(w^2 - N)\prod_\mu \Theta\left(\min_{x\in M^\mu} y^\mu\langle w, x\rangle - \kappa\right) \tag{4}$$

This is done using the replica method, which relies on the identity: $\mathbb{E}\log Z = \lim_{n\to 0} n^{-1}(\mathbb{E}Z^n - 1) = \lim_{n\to 0} n^{-1}\log\mathbb{E}Z^n$. We first assume that $n \in \mathbb{N}$ and only later take the limit $n \to 0$ after obtaining an expression which is analytic in n. Replicating the volume integral n times and rewriting the constraint in terms of the fields $H_i^{\mu,a}$ gives:

$$\mathbb{E}_{y,u}Z^n = \mathbb{E}_{y,u}\int\prod_{a=1}^n d^N w_a\delta(w_a^2 - N)\prod_\mu\int\mathbb{D}H^{\mu,a}\prod_{i=0}^K\sqrt{2\pi}\delta(H_i^{\mu,a} - y^\mu w_a^T u_i^\mu), \tag{5}$$

where, as in [2], we have absorbed the constraint into the measure $\mathbb{D}H$:

$$\mathbb{D}H^\mu = \left(\prod_{i=0}^K \frac{dH_i^\mu}{\sqrt{2\pi}}\right)\Theta\left(g_{\mathcal{S}^\mu}(H^\mu) - \kappa\right) \tag{6}$$

$$g_{\mathcal{S}^\mu}(H^\mu) = H_0^\mu + \min_{s\in\mathcal{S}^\mu}\sum_{i>0} H_i^\mu s_i \tag{7}$$

Introducing Fourier representations of the delta functions for H gives:

$$\int \prod_a d^N w_a \delta(w_a^2 - N) \Big(\prod_\mu \int \mathbb{D}H^{\mu,a} \prod_{i=0}^{K} \int \frac{d\hat{H}_i^{\mu,a}}{\sqrt{2\pi}} \Big) \mathbb{E}_{y,u} \exp\Big[\sum_{\mu,a,i} i\hat{H}_i^{\mu,a} (H_i^{\mu,a} - y^\mu w_a^T u_i^\mu) \Big] \tag{8}$$

The average over the exponential term is:

$$\mathbb{E}_{y,u} \exp\Big\{ - \sum_{\mu,a,i,l} i\hat{H}_i^{\mu,a} y^\mu w_a^l u_{i,l}^\mu \Big\} \tag{9}$$

$$= \mathbb{E}_y \int \frac{[\prod_{il\mu} du_{il}^\mu]}{(2\pi)^{NP(K+1)/2} |C|^{N/2}} \exp\Big\{ - \frac{N}{2} \sum_{\mu,i,\nu,j,l} u_{i,l}^\mu u_{j,l}^\nu (C^{-1})_{\nu,j}^{\mu,i} - i \sum_{\mu,a,i,l} \hat{H}_i^{\mu,a} y^\mu w_a^l u_{i,l}^\mu \Big\} \tag{10}$$

Note that the average over the labels y cannot be performed analytically. As such, the average $\mathbb{E}_y$ will not be written again until the end to avoid clutter. Defining the Cholesky decomposition of C, which satisfies $\sum_{\tau,k} L_{\tau,k}^{\mu,i} L_{\tau,k}^{\nu,j} = C_{\nu,j}^{\mu,i}$, we make the change of variables $u_{i,l}^\mu \mapsto \sum_{\tau,k} L_{\tau,k}^{\mu,i} u_{k,l}^\tau$. This gives:

$$\int \frac{[\prod_{il\mu} du_{il}^\mu]}{(2\pi)^{NP(K+1)/2}} \exp\Big\{ - \frac{N}{2} \sum_{\mu,i,l} \big(u_{i,l}^\mu\big)^2 - i \sum_{\mu,a,i,l,\tau,k} \hat{H}_i^{\mu,a} y^\mu w_a^l L_{\tau,k}^{\mu,i} u_{k,l}^\tau \Big\} \tag{11}$$

Integrating the u and introducing the overlap matrix $Q_{a,b} = N^{-1} \sum_{l=1}^N w_a^l w_b^l$ then yields:

$$\begin{aligned} &\int dQ \prod_a \int dw_a \delta(Q_{a,a} - 1) \prod_{b=1}^{n} \delta(N^{-1} w_a^T w_b - Q_{a,b}) \\ &\times \Big(\prod_\mu \int \mathbb{D}H^{\mu,a} \prod_{i=0}^{K} \int \frac{d\hat{H}_i^{\mu,a}}{\sqrt{2\pi}} \Big) \exp\Big\{ - \frac{1}{2} \sum_{a,b,\nu,\mu,i,j} Q_{a,b} C_{\nu,j}^{\mu,i} \hat{H}_i^{\mu,a} \hat{H}_j^{\nu,b} y^\mu y^\nu + i \sum_{a,\mu,i} \hat{H}_i^{\mu,a} H_i^{\mu,a} \Big\} \end{aligned} \tag{12}$$

Changing variables: $\hat{H}_i^{\mu,a} \mapsto y^\mu \sum_{\tau,k} [L^{-1}]_{\tau,k}^{\mu,i} \hat{H}_k^{\tau,a}$ and $H_i^{\mu,a} \mapsto y^\mu \sum_{\tau,k} L_{\tau,k}^{\mu,i} H_k^{\tau,a}$ gives:

$$\begin{aligned} &\int dQ \prod_a \int dw_a \delta(Q_{a,a} - 1) \prod_{b} \delta(N^{-1} w_a^T w_b - Q_{a,b}) \\ &\times \Big(\prod_a \int_{\mathcal{C}(y,L)} \Big[\prod_{\mu,i} dH_i^{\mu,a} \Big] \int \Big[\prod_{\mu,i} \frac{d\hat{H}_i^{\mu,a}}{\sqrt{2\pi}} \Big] \Big) \exp\Big\{ - \frac{1}{2} \sum_{a,b\mu,i} Q_{a,b} \hat{H}_i^{\mu,a} \hat{H}_i^{\mu,b} + i \sum_{a,\mu,i} \hat{H}_i^{\mu,a} H_i^{\mu,a} \Big\}, \end{aligned} \tag{13}$$

where the integral over each matrix $H^{,a}$ must now be taken over the set:

$$\mathcal{C}(y,L) \equiv \Big\{ H \in \mathbb{R}^{P \times (K+1)} : \forall \mu \in \mathbb{N}_1^p, \min_{s^\mu \in \mathcal{S}^\mu} y^\mu \sum_{k,\tau,i} H_k^\tau L_{\tau,k}^{\mu,i} s_i^\mu \geq \kappa \Big\} \tag{14}$$

Note that we have defined for convenience the additional element: $s_0^\mu = 1$, and that the sum is taken over: $0 \leq k, i \leq K$, and $1 \leq \mu, \tau \leq P$. Integrating the $\hat{H}$ variables:

$$\begin{aligned} &\int dQ \prod_a \int dw_a \delta(Q_{a,a} - 1) \prod_{b} \delta(N^{-1} w_a^T w_b - Q_{a,b}) \\ &\times \Big(\prod_a \int_{\mathcal{C}(y,L)} \Big[\prod_{\mu,i} dH_i^{\mu,a} \Big] \Big) \exp\Big\{ - \frac{1}{2} \sum_{a,b,\mu,i} Q_{a,b}^{-1} H_i^{\mu,a} H_i^{\mu,b} - \frac{P(K+1)}{2} \log\det Q \Big\} \end{aligned} \tag{15}$$

This is almost identical to the formula for the log volume for manifolds with heterogeneous shapes [2, 3]. The only difference is that the integration over the H is now constrained to the set $\mathcal{C}(y,L)$. As such, we can proceed just as in the case of uncorrelated manifolds with different shapes [2, 3], which we briefly describe here. We integrate over the w_a variables by introducing Fourier representations of the delta functions and carrying out the Gaussian integral over the w. The auxiliary variables introduced through this process can then be

integrated by saddle point, leading to a contribution $\exp\left[\frac{N}{2}\log\det Q + \text{const.}\right]$. From here, we assume replica symmetry: $Q_{a,b} = (1-q)\delta_{a,b} + q$ and apply a Hubbard-Stratonovich transformation to the off diagonal term in the summation over $H_i^{\mu,a}H_i^{\mu,b}Q_{a,b}^{-1}$ to arrive at:

$$\int \left[\prod_{a\neq b} dQ_{a,b}\right] \exp\left\{\frac{N-P(K+1)}{2}\log\det Q\right\}$$
$$\times \int D_I T\left[\int_{\mathcal{C}(y,L)}\left(\prod_{\mu,i} dH_i^\mu\right)\exp\left\{-\frac{1}{2(1-q)}||H-\sqrt{q}T||_2^2 + \frac{q}{2(1-q)}||T||_2^2\right\}\right]^n, \tag{16}$$

where just as in the main text, $D_I T$ denotes the isotropic Gaussian measure: $\prod_{\mu,i} dT_i^\mu \exp[-\frac{1}{2}(T_i^\mu)^2]/\sqrt{2\pi}$, and we have dropped the constant that emerged from the integral over the w. We now apply the identities: $\overline{f(x)^n} \doteq \exp[n\overline{\log f(x)}]$ and $\log\det Q \doteq n\log q + nq/(1-q)$, both of which hold as $n\to 0$. This gives the expression:

$$\int dq \exp\left\{\frac{Nn}{2}\left(\frac{q}{1-q} + \big(1-\alpha(K+1)\big)\log(1-q)\right) + \int D_I T \log\int_{\mathcal{C}(y,L)}\left(\prod_{\mu,i} dH_i^\mu\right) e^{-||H-\sqrt{q}T||_2^2/(2(1-q))}\right\} \tag{17}$$

We can see that q will concentrate around its saddle point value in the large N, P limit. This value of q is fixed by the stationarity condition:

$$\frac{1}{(1-q)^2} - \frac{1-\alpha(K+1)}{1-q} - \frac{1}{N(1-q)^2}\int D_I T \frac{\displaystyle\int_{\mathcal{C}(y,L)}\left(\prod_{\mu,i} dH_i^\mu\right)||H-T||_2^2 e^{-||H-\sqrt{q}T||_2^2/(2(1-q))}}{\displaystyle\int_{\mathcal{C}(y,L)}\left(\prod_{\mu,i} dH_i^\mu\right) e^{-||H-\sqrt{q}T||_2^2/(2(1-q))}} = 0 \tag{18}$$

Placing ourselves in the regime where we are at capacity requires that the volume of solutions shrinks to a single point. In this regime, the overlap between different solutions, q, concentrates about 1 [1]. Therefore, we can use the above self-consistency condition to determine the value of α such that exactly one solution to the separation problem will exist with probability 1 in the thermodynamic limit. To do this, we note that as $q\to 1$, the dominant terms are those of order $(1-q)^{-2}$, and the integrals over the H variables can be replaced by their values at the saddle point. Thus, to leading order in N, P, we can see that:

$$\frac{1}{\alpha_{corr}(\kappa)} = \mathbb{E}_y \int D_I T \min_{V\in\mathcal{C}(y,L)} \frac{1}{P}\sum_{\mu}^{P} ||V^\mu - T^\mu||_2^2, \tag{19}$$

where we have reintroduced the expectation over the labels and have switched from using H to V to match the main text. As described below, we use this form of the inverse capacity for all numerical calculations. Changing variables once more: $V_k^\tau \mapsto \sum_{\eta,l}[L^{-1}]_{\eta,l}^{\tau,k} y^\eta V_l^\eta$, and $T_k^\tau \mapsto \sum_{\eta,l}[L^{-1}]_{\eta,l}^{\tau,k} y^\eta T_l^\eta$ then gives the stated result:

$$\mathbb{E}_y \int D_{y,C} T \min_{V\in\mathcal{A}} \frac{1}{P}\sum_{\mu,\nu,i,j}(V_i^\mu - T_i^\mu)(V_j^\nu - T_j^\nu)\big(C^{-1}\big)_{\nu,j}^{\mu,i} y^\mu y^\nu \tag{20}$$

$$\mathcal{A} = \left\{V\in\mathbb{R}^{P\times(K+1)} : \forall\mu\in\mathbb{N}_1^P,\ V_0^\mu + \min_{s\in\mathcal{S}^\mu}\sum_{i>0} V_i^\mu s_i \geq \kappa\right\} \tag{21}$$

$$D_{y,C}T \equiv (2\pi)^{-P(K+1)/2}|C|^{-1/2}\exp\left\{-\frac{1}{2}\sum_{\mu,\nu,i,j} T_i^\mu T_j^\nu y^\mu y^\nu \big(C^{-1}\big)_{\nu,j}^{\mu,i}\right\}\left[\prod_{\mu=1}^{P}\prod_{i=0}^{K} dT_i^\mu\right], \tag{22}$$

where we have used the fact that the determinant of $y^\mu y^\nu C_{\nu,j}^{\mu,i}$ is unaffected by the off-diagonal sign flips $y^\mu y^\nu$ to write the normalizing constant over the T integral as $(2\pi)^{-P(K+1)/2}|C|^{-1/2}$. (This can be derived by considering the matrix integral $\int(\prod_{\mu,i} dT_i^\mu)\exp[-\frac{1}{2}y^\mu y^\nu T_i^\mu T_j^\nu (C^{-1})_{\nu,j}^{\mu,i}]$ and changing variables $T_i^\mu \mapsto y^\mu T_i^\mu$.)

□

2 The Capacity for Spheres with Low-Rank Correlations

Claim 2. *Consider a set of P spheres of radius r, intrinsic dimension K, and at a distance r_0 from the origin, residing in $\mathbb{R}^N$. Given homogenous axis-axis and centroid-centroid correlations as defined in the main text Eq. 6, the capacity for these spheres is given by:*

$$\frac{1}{\alpha_{corr}(\kappa)} = K\left(\sqrt{q}-1\right)^2 + \int_{-\infty}^{\hat{\kappa}(q)} \frac{d\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \left(\xi - \hat{\kappa}(q)\right)^2 , \tag{23}$$

where the scaled squared norm of the signed fields $q \equiv \overline{\sum_{i>0}(V_i^\mu)^2}/((1-\lambda)K)$ and the effective margin $\hat{\kappa}(q)$ are fixed by the equations:

$$\sqrt{q} = 1 + \frac{r\sqrt{1-\lambda}}{r_0\sqrt{K(1-\psi)}} \int_{-\infty}^{\hat{\kappa}(q)} \frac{d\xi}{\sqrt{2\pi}} e^{-\frac{1}{2}\xi^2} \left(\xi - \hat{\kappa}(q)\right) \tag{24}$$

$$\hat{\kappa}(q) = \frac{r\sqrt{K(1-\lambda)q} + \kappa}{r_0\sqrt{1-\psi}} \tag{25}$$

Derivation: To start with, we rewrite the formula for the inverse capacity given in Supplementary Eq. 1 in terms of a Gibbs measure [4]:

$$\frac{1}{\alpha_{cor}(\kappa)} = \lim_{\beta\to\infty}\lim_{P\to\infty} -\frac{2}{P}\frac{\partial}{\partial\beta}\mathbb{E}_{T,y}\log\int\left[\prod_{\mu,i} dV_i^\mu\right]\exp\left[-\frac{\beta}{2}y^\mu y^\nu \Lambda^{\mu,i}_{\nu,j}(V_i^\mu - T_i^\mu)(V_j^\nu - T_j^\nu)\right] \times \prod_\mu \Theta\left(r_0 V_0^\mu - \kappa + \min_{s\in\mathcal{S}^\mu}\sum_{i>0} V_i^\mu s_i r\right) , \tag{26}$$

where $\Lambda^{\mu,i}_{\nu,j}$ is the inverse covariance tensor, $(C^{-1})^{\mu,i}_{\nu,j}$. In this form, we can see that the capacity can be derived from the disorder averaged free energy density, $-(\beta P)^{-1}\mathbb{E}_{T,y}\log Z$. We calculate this using the replica method [5, 6]. Here and in the remainder of the calculation, we implicitly sum over all indices in the exponent unless noted otherwise. That is: $\exp[f(x_{a,b})] \equiv \exp[\sum_{a,b} f(x_{a,b})]$. By the Sherman-Morrison formula [7], $\Lambda^{\mu,i}_{\nu,j}$ has entries:

$$\Lambda^{\mu,i}_{\nu,j} = \begin{cases} \delta_{ij}\left[\dfrac{\delta_{\mu,\nu}}{1-\lambda} - \dfrac{\lambda}{(1-\lambda)(1+(P-1)\lambda)}\right] & \text{for } i>0 \\ \dfrac{\delta_{\mu,\nu}}{1-\psi} - \dfrac{\psi}{(1-\psi)(1+(P-1)\psi)} & \text{for } i=j=0 \\ 0 & \text{for } i=0, j\neq 0 \end{cases} \tag{27}$$

$$\doteq \begin{cases} \delta_{ij}\left[\dfrac{\delta_{\mu,\nu}}{1-\lambda} - \dfrac{1}{P(1-\lambda)}\right] & \text{for } i>0 \\ \dfrac{\delta_{\mu,\nu}}{1-\psi} - \dfrac{1}{P(1-\psi)} & \text{for } i=j=0 \\ 0 & \text{for } i=0, j\neq 0 \, , \end{cases} \tag{28}$$

where $\doteq$ denotes equality to leading order in P. Using the assumption of spherical manifolds allows us to carry out the constrained minimization in the Θ functions above:

$$r_0 V_0^\mu + \min_{s\in\mathcal{S}^\mu}\sum_{i>0} r V_i^\mu s_i = r_0 V_0^\mu - r\sqrt{\sum_{i>0}\left(V_i^\mu\right)^2} \tag{29}$$

To see this, we apply the KKT conditions to the Lagrangian: $L(s,\eta) = \langle V, s\rangle + \eta(||s||^2 - 1)$ [8]. The KKT conditions read:

$$2\eta s = -V \tag{30}$$

$$\eta \geq 0 \tag{31}$$

$$||s||^2 \leq 1 \tag{32}$$

$$\eta(||s||^2 - 1) = 0 \tag{33}$$

For all non-zero V, we must therefore have $\eta > 0$, from which it follows that $s = -V/||V||$ at the minimum, establishing the identity in Eq. (29). Using this simplification and replicating the partition function above n times then gives:

$$\mathbb{E}_{T,y} Z^n = \mathbb{E}_{T,y} \int \Big[\prod_{\mu,i,a} dV_{a,i}^{\mu} \Big] \exp \Big[-\frac{\beta}{2} y^{\mu} y^{\nu} \Lambda_{\nu,j}^{\mu,i} (V_{a,i}^{\mu} - T_i^{\mu})(V_{a,j}^{\nu} - T_j^{\nu}) \Big] \prod_{a,\mu} \Theta\Big(V_{a,0}^{\mu} - \frac{r}{r_0} ||V_{a,i>0}^{\mu}|| - \frac{\kappa}{r}_0 \Big) , \tag{34}$$

where we have used $||V_{i>0}||$ to denote the norm of the last K components of V_a^{μ}. That is, $||V_{a,i>0}^{\mu}|| \equiv \sqrt{\sum_{i>0} \left(V_{a,i}^{\mu}\right)^2}$. The expectation over T is:

$$\int d^{P\times(K+1)} T \exp \Big[-\frac{1+\beta n}{2} T_i^{\mu} T_j^{\nu} y^{\mu} y^{\nu} \Lambda_{\nu,j}^{\mu,i} + \beta T_i^{\mu} y^{\mu} \Lambda_{\nu,j}^{\mu,i} y^{\nu} V_{a,j}^{\nu} - \frac{1}{2} \log\det yy^T \circ \Lambda - \frac{P(K+1)}{2} \log 2\pi \Big] \tag{35}$$

$$= \exp \Big[\frac{\beta^2}{2(1+\beta n)} V_{a,i}^{\mu} V_{b,j}^{\nu} y^{\mu} y^{\nu} \Lambda_{\nu,j}^{\mu,i} - \frac{P(K+1)}{2} \log(1+n\beta) \Big] \tag{36}$$

Here we have slightly abused notation to write: $(yy^T \circ \Lambda)_{\nu,j}^{\mu,i} = y^{\mu} y^{\nu} \Lambda_{\nu,j}^{\mu,i}$. Note that the replicas V_a are now coupled after integrating out the quenched disorder T. Reinserting this back into the integral and expanding the terms $\log(1+n\beta)$, Λ, and $1/2(1+\beta n)$ to first order in the $n \to 0$ and $P \to \infty$ limits:

$$\begin{aligned} \mathbb{E}_{T,y} Z^n \doteq \mathbb{E}_y \int \Big[\prod_{\mu,a} \prod_{i>0} dV_{a,i}^{\mu} \Big] \exp \Big[-\sum_{i>0} \frac{\beta}{2(1-\lambda)} \Big\{ \left(V_{a,i}^{\mu}\right)^2 - \frac{1}{P} \Big(\sum_{\mu} y^{\mu} V_{a,i}^{\mu} \Big)^2 \Big\} \\ + \frac{\beta^2}{2(1-\lambda)} \Big\{ V_{a,i}^{\mu} V_{b,i}^{\mu} - \frac{1}{P} \Big(\sum_{a,\mu} y^{\mu} V_{a,i}^{\mu} \Big)^2 \Big\} - \frac{\beta P(K+1) n}{2} \Big] \\ \times \int \Big[\prod_{\mu,a} dV_{a,0}^{\mu} \Big] \exp \Big[-\frac{\beta}{2(1-\psi)} \Big\{ \left(V_{a,0}^{\mu}\right)^2 - \frac{1}{P} \Big(\sum_{\mu} y^{\mu} V_{a,i}^{\mu} \Big)^2 \Big\} \\ + \frac{\beta^2}{2(1-\psi)} \Big\{ V_{a,i}^{\mu} V_{b,i}^{\mu} - \frac{1}{P} \Big(\sum_{a,\mu} y^{\mu} V_{a,0}^{\mu} \Big)^2 \Big\} \Big] \prod_{a,\mu} \Theta\Big(V_{a,0}^{\mu} - \frac{r}{r_0} ||V_{i>0}|| - \frac{\kappa}{r}_0 \Big) \end{aligned} \tag{37}$$

Note that we freely ignore constants which do not affect the final result. The important points here are: (1) the only interaction between manifolds happens through the mean, $\sum_{\mu} y^{\mu} V_{a,i}^{\mu}$, and (2) the only interactions between the replicas happens through the quadratic interaction terms, $V_{a,i}^{\mu} V_{b,i}^{\mu}$. These considerations motivate the following substitutions, which we enforce using delta functions:

$$Q_{a,b}^{\mu} = \frac{1}{K} \sum_{i>0} V_{a,i}^{\mu} V_{b,i}^{\mu} \tag{38}$$

$$F_{a,i} = \frac{1}{P} \sum_{\mu} y^{\mu} V_{a,i}^{\mu} \tag{39}$$

In this way, equation (37) becomes:

$$\mathbb{E}_y \int \Big[\prod_{a \le b} dQ_{a,b} \Big] \exp \Big[-\frac{K\beta}{2(1-\lambda)} (Q_{a,a}^{\mu} - \beta Q_{a,b}^{\mu}) - \frac{P(K+1)\beta n}{2} + \log \mathcal{S}(Q) + \log \mathcal{U}(Q) \Big], \tag{40}$$

where we have defined for convenience $Q_{a,b} \equiv Q_{b,a}$, and $\mathcal{U}$ and $\mathcal{S}$ are the remaining integrals over the $V_{a,i}^{\mu}$ and the F:

$$\mathcal{S}(Q) \equiv \int \Big[\prod_{a,i>0} dF_{a,i} \Big] \exp \Big[\frac{\beta}{2(1-\lambda)} \Big\{ P F_{a,i}^2 - \beta P \Big(\sum_{a} F_{a,i} \Big)^2 \Big\} \Big] \int \Big[\prod_{a,\mu} \prod_{i>0} dV_{a,i}^{\mu} \Big] \tag{41}$$

$$\times \Big[\prod_{\mu} \prod_{a \le b} \delta\Big(Q_{a,b}^{\mu} - K^{-1} \sum_{i>0} V_{a,i}^{\mu} V_{b,i}^{\mu} \Big) \Big] \prod_{i,a} \delta\Big(F_{a,i} - P^{-1} \sum_{\mu} y^{\mu} V_{a,i}^{\mu} \Big) \tag{42}$$

$$\mathcal{U}(Q) \equiv \int \Big[\prod_a F_{a,0}\Big] \exp\Big[\frac{\beta}{2(1-\psi)}\Big\{PF_{a,0}^2 - \beta P\Big(\sum_a F_{a,0}\Big)^2\Big\}\Big] \tag{43}$$

$$\times \int \Big[\prod_{a,\mu} dV_{a,0}^{\mu}\Big]\Big[\prod_i \delta\Big(F_{a,0} - P^{-1}\sum_\mu y^\mu V_{a,0}^\mu\Big)\Big]\prod_{a,\mu}\Theta\Big(V_{a,0}^\mu - R\sqrt{Q_{a,a}} - \kappa r_0^{-1}\Big)$$

$$\times \exp\Big[-\frac{\beta}{2(1-\psi)}\big(V_{a,0}^\mu\big)^2 + \frac{\beta^2}{2(1-\psi)}V_{a,0}^\mu V_{b,0}^\mu\Big] \tag{44}$$

Note that we have defined $R \equiv \sqrt{K}r/r_0$.
In order to evaluate $\mathcal{U}(Q)$ and $\mathcal{S}(Q)$ in closed form, we now make the replica symmetric ansatz:

$$Q_{a,b}^\mu = \delta_{a,b}(q_0 - q_1) + q_1 \tag{45}$$

Note that unlike typical assumptions of replica symmetric ansatz [5, 6], we have to assume symmetry across all manifolds: $Q^\mu = Q$. This assumption is motivated by the fact that the interactions between all manifolds are symmetric. Under this assumption, the function $S(Q)$ can be estimated by saddle point as described in lemma 3. In this way we obtain in the $n \to 0$ limit:

$$\int dq_0 dq_1 \exp\Big[-\frac{nPK\beta}{2(1-\lambda)}\big(q_0 - \beta(q_0 - q_1)\big) - \frac{P(K+1)\beta n}{2} + \frac{1}{2}PK\log\det Q + \log\mathbb{E}_y\mathcal{U}\big(Q\big)\Big], \tag{46}$$

where in the $n \to 0$ limit we have that:

$$\log\det Q = n\log(q_0 - q_1) + n\frac{q_1}{q_0 - q_1} \tag{47}$$

In lemma 4, we evaluate the $\mathbb{E}_y\mathcal{U}(Q)$ by saddle point. Plugging this result into the integral gives:

$$\int dq_0 dq_1 \exp\Big[\frac{PKn}{2}\Big\{-\frac{\beta}{1-\lambda}\big(q_0 - \beta(q_0 - q_1)\big) - \beta + \log(q_0 - q_1) + \frac{q_1}{q_0 - q_1}\Big\} \tag{48}$$

$$+ nP\int \mathcal{D}_\psi\xi \log\mathcal{H}\Big(\sqrt{\frac{\beta}{1-\psi}}\big(R\sqrt{q_0} + \frac{\kappa}{r}_0 - \xi\big)\Big)\Big],$$

where, as below, we have used $\mathcal{H}$ to denote the unnormalized Gaussian tail function: $\mathcal{H}(x) \equiv \int_x^\infty dt\exp(-t^2/2)$ and $\mathcal{D}_\psi\xi$ to denote the zero-mean Gaussian measure with variance $1-\psi$. Pulling out the P, n factors, we can see that the replicated partition can be written as:

$$\mathbb{E}_{T,y}Z^n \doteq \int dq_0 dq_1 e^{nP\mathcal{V}(q_0,q_1)} \tag{49}$$

If we estimate the remaining integral by saddle point in the large P limit, we can use the identity $\overline{\log Z} = \lim_{n\to 0} n^{-1}\log\overline{Z^n}$ to arrive at:

$$\frac{1}{P}\mathbb{E}_{T,y}\log Z \doteq \underset{q_0,q_1}{\text{extr}}\,\mathcal{V}(q_0, q_1) \tag{50}$$

We therefore have to solve $\nabla V(q_0, q_1) = 0$. Just as in lemma 4 we can use the expansion of $\mathcal{H}(\sqrt{\beta}x)$ as $\beta \to \infty$ given in equation 71 to see that:

$$\partial_{q_0}\int \mathcal{D}_\psi\xi\log\mathcal{H}\Big(\sqrt{\frac{\beta}{1-\psi}}\big(R\sqrt{q_0} + \kappa r_0^{-1} - \xi\big)\Big) = -\frac{R\sqrt{\beta}}{2\sqrt{q_0(1-\psi)}}\int \mathcal{D}_\psi\xi\frac{\exp\Big[-\frac{\beta(R\sqrt{q_0}+\kappa r_0^{-1}-\xi)^2}{2(1-\psi)}\Big]}{\mathcal{H}\Big(\sqrt{\frac{\beta}{1-\psi}}\big(R\sqrt{q_0} + \kappa r_0^{-1} - \xi\big)\Big)}$$

$$\doteq -\frac{R\beta}{2(1-\psi)\sqrt{q_0}}\int_{-\infty}^{R\sqrt{q_0}+\kappa r_0^{-1}} \mathcal{D}_\psi\xi(R\sqrt{q_0} - \kappa r_0^{-1} - \xi) \tag{51}$$

Using this expansion, the function $\mathcal{V}$ admits two different pairs of saddle points. The meaningful solution is given by:

$$q_1 = q_0 - \frac{1}{\beta}\sqrt{q_0(1-\lambda)} + O(\beta^{-2}) \tag{52}$$

$$\sqrt{q_0} = \sqrt{1-\lambda} + \frac{r(1-\lambda)}{r_0\sqrt{K(1-\psi)}}\int_{-\infty}^{0}\frac{d\xi}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\xi+\frac{r\sqrt{Kq_0}+\kappa}{r_0\sqrt{1-\psi}}\right)^2}\xi \tag{53}$$

Denoting the solutions to the above equations as q_0^*, q_1^*, the inverse capacity is then given by

$$-2\frac{\partial}{\partial\beta}\mathcal{V}(q_0^*, q_1^*) = K\left(\sqrt{\frac{q_0^*}{1-\lambda}} - 1\right)^2 + \int_{-\infty}^{0}\frac{d\xi}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\xi+\frac{r\sqrt{Kq_0^*}+\kappa}{r_0\sqrt{1-\psi}}\right)^2}\xi^2, \tag{54}$$

where again we have used the same expansion of $\mathcal{H}$ to evaluate the partial derivative with respect to β. From here, changing $q_0 \mapsto q_0(1-\lambda)$ then gives the stated result.

□

Lemma 3. *Under the replica symmetric ansatz and as $\beta, P \to \infty$ and $n \to 0$, the function $S(Q)$ is asymptotic to:*

$$\exp\left[\frac{1}{2}PK\log\det Q + \text{const.}\right], \tag{55}$$

where const. *denotes terms which do not depend on β or Q.*

Derivation: Introducing Fourier representations of the delta functions in equation 42 gives:

$$\int\Big[\prod_{\mu,a\leq b}d\hat{Q}^{\mu}_{a,b}\Big]\int\Big[\prod_{i>0,a}dF_{a,i}d\hat{F}_{a,i}\Big]\exp\left[\frac{i}{2}K\sum_{a\leq b}Q_{a,b}\hat{Q}^{\mu}_{a,b} + \sum_{a,i}\frac{\beta}{2(1-\lambda)}\left\{PF_{a,i}^2 - \beta P\Big(\sum_a F_{a,i}\Big)^2\right\}\right]$$
$$\times\exp\Big[iPF_{a,i}\hat{F}_{a,i}\Big]\int\Big[\prod_{a,\mu}\prod_{i>0}dV^{\mu}_{a,i}\Big]\exp\Big[-\frac{i}{2}\sum_{a\leq b}\hat{Q}^{\mu}_{a,b}V^{\mu}_{a,i}V^{\mu}_{b,i} - iy^{\mu}\hat{F}_{a,i}V^{\mu}_{a,i}\Big] \tag{56}$$

It is convenient to now rewrite the ordered sum over $a \leq b$ indices as an unordered sum over all pairings (a, b). As above, we define $Q_{b,a} \equiv Q_{a,b}$ and $\hat{Q}_{b,a} \equiv Q_{a,b}$ for $a > b$. If we further change $\hat{Q}_{a,b} \mapsto 2\hat{Q}_{a,b}$ for all $a \neq b$, we can eliminate the unordered sums as desired, leaving:

$$\int\Big[\prod_{\mu,a\leq b}d\hat{Q}^{\mu}_{a,b}\Big]\int\Big[\prod_{i>0,a}dF_{a,i}d\hat{F}_{a,i}\Big]\exp\left[\frac{i}{2}KQ_{a,b}\hat{Q}^{\mu}_{a,b} + \frac{\beta}{2(1-\lambda)}\left\{PF_{a,i}^2 - \beta P\Big(\sum_a F_{a,i}\Big)^2\right\}\right]$$
$$\times\exp\Big[iPF_{a,i}\hat{F}_{a,i}\Big]\int\Big[\prod_{a,\mu}\prod_{i>0}dV^{\mu}_{a,i}\Big]\exp\Big[-\frac{i}{2}\hat{Q}^{\mu}_{a,b}V^{\mu}_{a,i}V^{\mu}_{b,i} - iy^{\mu}\hat{F}_{a,i}V^{\mu}_{a,i}\Big], \tag{57}$$

where as usual we neglect constants which do not depend on either the variables of integration or β, as they will not affect the final result. The integrals over the V followed by the $\hat{F}$ variables are now both standard Gaussian integrals. They yield:

$$\int\Big[\prod_{\mu,a\leq b}d\hat{Q}^{\mu}_{a,b}\Big]\Big[\prod_{a}\prod_{i>0}dF_{a,i}\Big]\exp\left[\frac{i}{2}KQ_{a,b}\hat{Q}^{\mu}_{a,b} - \frac{K}{2}\sum_{\mu}\log\det\hat{Q}^{\mu} - \frac{1}{2}\log\det\Big(\sum_{\mu}\left(\hat{Q}^{\mu}\right)^{-1}\Big)\right.$$
$$\left.-\frac{P^2}{2}F_{a,i}F_{b,i}\Big(\sum_{\mu}(\hat{Q}^{\mu})^{-1}\Big)^{-1}_{a,b} + \frac{\beta}{2(1-\lambda)}\left\{PF_{a,i}^2 - \beta P\Big(\sum_a F_{a,i}\Big)^2\right\}\right] \tag{58}$$

The remaining Gaussian integral over the F produces a term which is subleading in P. Therefore, the leading order in the remaining integral over the $\hat{Q}$ is simply the first two terms in the exponent. If we now invoke the replica symmetric ansatz on the conjugate variables: $\hat{Q}^{\mu} = \hat{Q}$, we are left with:

$$\int\Big[\prod_{a\leq b}\hat{Q}_{a,b}\Big]\exp\left[\frac{i}{2}PKQ_{a,b}\hat{Q}_{a,b} - \frac{PK}{2}\log\det\hat{Q}\right] \tag{59}$$

Estimating this integral by saddle point then gives the desired result. □

Lemma 4. *Under the replica symmetric ansatz, as $\beta, P \to \infty$ and $n \to 0$, the function $\mathcal{U}(Q)$ is asymptotic to:*

$$\mathcal{U}(Q) \doteq \exp\left[nP \int \mathcal{D}_\psi \xi \log \mathcal{H}\Big(\sqrt{\frac{\beta}{1-\psi}}\big(R\sqrt{q_0} + \frac{\kappa}{r_0} - \xi\big)\Big) + \frac{1}{2}\beta P n + \text{const.}\right], \tag{60}$$

where, as above, const. *denotes terms which do not depend on β or Q, and $\mathcal{H}$ is the unnormalized Gaussian tail function:*

$$\mathcal{H}(x) \equiv \int_x^\infty ds e^{-s^2/2} \tag{61}$$

We also use $\mathcal{D}_\psi \xi$ to denote the Gaussian measure:

$$\mathcal{D}_\psi \xi \equiv \frac{d\xi e^{-\xi^2/2(1-\psi)}}{\sqrt{2\pi(1-\psi)}} \tag{62}$$

Derivation: Introducing Fourier representations of the delta functions in equation (44):

$$\mathbb{E}_y \int \Big[\prod_a dF_{a,0} d\hat{F}_a \prod_\mu dV^\mu_{a,0}\Big] \exp\Big[-\frac{\beta}{2(1-\psi)}\big(V^\mu_{a,0}\big)^2 + \frac{\beta^2}{2(1-\psi)}\Big(\sum_a V^\mu_{a,0}\Big)^2\Big] \prod_{a,\mu} \Theta\Big(V^\mu_{a,0} - R\sqrt{q_0} - \frac{\kappa}{r_0}\Big)$$
$$\times \exp\Big[\frac{P\beta}{2(1-\psi)}\Big\{F^2_{a,i} - \Big(\sum_a F_{a,i}\Big)^2\Big\} - \frac{i\beta}{1-\psi}\hat{F}_a\big(PF_{a,0} - y^\mu V^\mu_{a,0}\big) + O(\log\beta)\Big] \tag{63}$$

Note that the terms $O(\log\beta)$ will have no effect on our final answer, so we ignore them. If we now make the additional replica symmetric assumption: $\hat{F}_a = \hat{F}, F_a = F$ and introduce a Hubbard-Stratonovich transform on each of the terms $(\sum_a V^\mu_{a,0})^2$, we have:

$$\int dF d\hat{F} \exp\Big[\frac{P\beta n}{2(1-\psi)}F^2 - \frac{iPn\beta}{1-\psi}F\hat{F} + O(n^2)\Big] \prod_\mu \mathbb{E}_{y^\mu} \int \mathcal{D}_\psi \xi_\mu \tag{64}$$
$$\times \Big[\int dV^\mu_0 \exp\Big[-\frac{\beta}{2(1-\psi)}\big(V^\mu_0\big)^2 + \frac{\beta}{1-\psi}V^\mu_0(\xi_\mu + i\hat{F}y^\mu)\Big]\Theta\Big(V^\mu_0 - R\sqrt{q_0} - \frac{\kappa}{r_0}\Big)\Big]^n \tag{65}$$

We can factorize the integral across the μ-index and absorb the Θ function into the limits of integration of the V_0 variables to obtain:

$$\int dF d\hat{F}\Big[\int \mathcal{D}_\psi \xi \mathbb{E}_y\Big(\int_{R\sqrt{q_0}+\kappa}^\infty dV_0 \exp\Big[-\frac{\beta}{2(1-\psi)}(V_0 - \xi - i\hat{F}y)^2 + \frac{\beta}{2(1-\psi)}(\xi + i\hat{F}y)^2\Big]\Big)^n\Big]^P$$
$$\times \exp\Big[-\frac{iPn\beta F\hat{F}}{1-\psi} + \frac{P\beta n}{2(1-\psi)}F^2\Big] \tag{66}$$

Using the identity $\mathbb{E}_x f(x)^n \doteq e^{n\mathbb{E}_x \log f(x)}$, which is valid in the $n \to 0$ limit, we obtain, after changing variables $V \mapsto V + \xi + i\hat{F}y$:

$$\int dF d\hat{F} \exp\Big[nP\mathbb{E}_y \int \mathcal{D}_\psi \xi \log\Big(\int_{R\sqrt{q}-\xi-i\hat{F}y}^\infty e^{-\beta V_0^2/2(1-\psi)}\Big) + nP\mathbb{E}_y \int \mathcal{D}_\psi \xi \frac{\beta}{2(1-\psi)}(\xi + i\hat{F}y)^2\Big] \tag{67}$$
$$\times \exp\Big[-\frac{iPn\beta F\hat{F}}{1-\psi} + \frac{P\beta n}{2(1-\psi)}F^2\Big] \tag{68}$$

Carrying out the expectations over the terms $(\xi + i\hat{F}y)^2$ and changing $V \mapsto \sqrt{(1-\psi)/\beta}V$ then gives:

$$\int dF d\hat{F} \exp\Big[nP\mathbb{E}_y \int \mathcal{D}_\psi \xi \log \mathcal{H}\Big(\sqrt{\frac{\beta}{1-\psi}}\big(R\sqrt{q_0} + \kappa r_0^{-1} - \xi - i\hat{F}y\big)\Big)$$
$$+ \frac{1}{2}\beta P n - \frac{nP\beta}{2(1-\psi)}\hat{F}^2 - \frac{iPn\beta F\hat{F}}{1-\psi} + \frac{P\beta n}{2(1-\psi)}F^2 + O(nP\log\beta)\Big] \tag{69}$$

We are now ready to estimate this integral by saddle point. To do so, we start by noting that the partial derivative with respect to $\hat{F}$ of the average over the $\log \mathcal{H}$ term is:

$$\sqrt{\frac{\beta}{1-\psi}}\mathbb{E}_y \int \mathcal{D}_\psi \xi \frac{yi \exp\left[-\frac{\beta}{2(1-\psi)}\left(R\sqrt{q_0} + \kappa r_0^{-1} - \xi - i\hat{F}y\right)^2\right]}{\mathcal{H}\left[\sqrt{\frac{\beta}{1-\psi}}\left(R\sqrt{q_0} + \kappa r_0^{-1} - \xi - i\hat{F}y^\mu\right)\right]} \tag{70}$$

In the $\beta \to \infty$ limit, we can use the expansion:

$$\mathcal{H}(\sqrt{\beta}x) \doteq \begin{cases} \frac{e^{-\beta x^2/2}}{\sqrt{\beta}x} & x > 0 \\ \sqrt{2\pi} & x < 0 \end{cases} \tag{71}$$

Invoking this expansion, the expression simplifies to:

$$\frac{i\beta}{1-\psi}\mathbb{E}_y \int_{-\infty}^{0} \frac{d\xi e^{-(\xi + R\sqrt{q_0} + \kappa r_0^{-1} - i\hat{F}y)^2/2(1-\psi)}}{\sqrt{2\pi(1-\psi)}} \xi y \tag{72}$$

The saddle points over the $F, \hat{F}$ variables then satisfy the self-consistent equations:

$$F = i\hat{F} \tag{73}$$

$$-iF - \hat{F} + i\mathbb{E}_y \int_{-\infty}^{0} \frac{d\xi e^{-(\xi + R\sqrt{q_0} + \kappa r_0^{-1} - i\hat{F}y)^2/2(1-\psi)}}{\sqrt{2\pi(1-\psi)}} \xi y = 0, \tag{74}$$

which have the solution $\hat{F} = F = 0$. Replacing equation (69) with its value at the saddle point, we can see that the function $\mathbb{E}_y U(Q)$ can be stated as:

$$\exp\left[nP \int \mathcal{D}_\psi \xi \log \mathcal{H}\left(\sqrt{\frac{\beta}{1-\psi}}\left(R\sqrt{q_0} + \kappa r_0^{-1} - \xi\right)\right) + \frac{1}{2}\beta P n\right], \tag{75}$$

which is what we wanted to show. □

Algorithm 1 Capacity Estimation for Correlated Data (α_{cor})

Input: n_t : Number of Monte Carlo draws. G : Data array of shape $P \times N \times M$ containing M samples of P distinct manifolds, in an ambient dimension of N.
Output: α : Capacity estimate.

$L, S \leftarrow \text{GetShapesAndCholesky}(G)$
$\alpha^{-1} \leftarrow \text{ZerosArray}(n_t)$
for i from 1 to n_t **do**
    $y \leftarrow Bernoulli^{\otimes P}(0.5)$
    $T \leftarrow Normal^{\otimes P \times M+1}(0, 1)$
    $\alpha^{-1}[i] \leftarrow \min_{V \in \mathcal{C}(y,L)} \frac{1}{P} \sum_{\mu}^{P} ||V^{\mu} - T^{\mu}||^2$
end for
return $1/\text{mean}(\alpha^{-1})$

Algorithm 2 GetShapesAndCholesky

Input: G : Data array of shape $P \times N \times M$ containing M samples of P distinct manifolds, in an ambient dimension of N.
Output: L : Cholesky decomposition of the correlation tensor, reshaped into a $P(M+1) \times P(M+1)$ matrix. S : An array of shape $P \times M \times M$ containing data points in their axis coordinates.

$Ax, S \leftarrow \text{ZerosArray}(P, N, M+1), \text{ZerosArray}(P, M, M)$
for μ from 1 to P **do**
    $c \leftarrow M^{-1} \sum_{i=1}^{M} G[\mu, :, i]$ ▷ Get this manifold's centroid
    for i from 1 to M **do**
        $G[\mu, :, i] \leftarrow G[\mu, :, i] - c$ ▷ Center each sample
    end for
    $Ax[\mu, :, 0] \leftarrow c$ ▷ Assign centroid to leading dimension
    $Ax[\mu, :, 1:], S[\mu] \leftarrow \text{QR}(G[\mu])$ ▷ Assign axes to the next m dimensions, and get the manifold points in the manifold axis coordinates, $\mathcal{S}^{\mu}$
end for
$Ax \leftarrow \text{reshape}(Ax, (P * (M+1), N))$
if $\text{Rank}(Ax) < P * (M+1)$ **then**
    $C \leftarrow AxAx^T + \epsilon I$ ▷ Perturb the diagonal by a small ϵ to make C positive definite (we set $\epsilon = 0.001$).
    $L \leftarrow \text{Cholesky}(C)$
else
    $Q, L^T \leftarrow \text{QR}(Ax)$ ▷ When full rank, use the QR decomposition to get L.
end if
return L, S

3 Numerical Implementation of the Capacity Estimator:

In this section, we describe how we estimate the capacity for arbitrary data manifolds. While there are several ways to parameterize the data manifolds in terms of axes and shape sets, we use the QR decomposition of the matrices containing the (centered) manifold data points to obtain a set of orthogonal axes vectors (Q), together with the coordinates of each manifold point in this orthogonal basis (R), as described in steps 1-8 of Algorithm 2. The shape sets $\mathcal{S}^\mu$ are then simply taken to be the manifold points in this basis. With respect to the quadratic minimization in Eqs (1) and (76) with linear constraint sets enforcing separability in Eqs. (2) and (77), this choice is equivalent to taking the shape sets to be the convex hull of all manifold points [8]. With these definitions in hand, the correlation tensor C can then simply be taken to be the empirical covariance tensor, $C^{\mu,i}_{\nu,j} = \langle u^\mu_i, u^\nu_j \rangle$.

The direct estimation of Supplementary Eq. (1) using the data manifolds parameterized as described above is a difficult problem. The main difficulty comes from the inversion of large correlation tensors, which is a highly numerically unstable operation. We sidestep this difficulty by changing variables: $(V-T)^\mu_i \mapsto \sum_{\tau,k} y^\mu L^{\mu,i}_{\tau,k}(V-T)^\tau_k$. Here L is the Cholesky factorization of the correlation tensor, C, which satisfies $\sum_{\tau,k} L^{\mu,i}_{\tau,k} L^{\nu,j}_{\tau,k} = C^{\mu,i}_{\nu,j}$. This change of variables gives the representation:

$$\frac{1}{\alpha_{cor}(\kappa)} = \overline{\int D_I T \min_{V \in \mathcal{C}(y,L)} \frac{1}{P} \sum_{\mu}^{P} ||V^\mu - T^\mu||^2} \tag{76}$$

$$\mathcal{C}(y,L) = \left\{ V \in \mathbb{R}^{P \times (K+1)} : \min_{s^\mu \in \mathcal{S}^\mu} y^\mu \sum_{k,\tau,i} V^\tau_k L^{\mu,i}_{\tau,k} s^\mu_i \geq \kappa \right\}, \tag{77}$$

where we have defined $s^\mu_0 \equiv 1$ for convenience. We can see that estimating Eq. (76) with Monte Carlo draws of y, T now only requires calculating the Cholesky factorization of the covariance. Even when N is very large, this step can safely be done using the QR decomposition of the matrix containing manifold axes and centroids (Algorithm 2, line 15). Note, however, that this step requires that the correlation tensor be full rank, which may not always be the case (e.g., when $N < MP$). Therefore, when the covariance tensor is not full rank, we add a small perturbation to the diagonal of the correlation tensor and calculate the Cholesky factor directly (Algorithm 2, lines 11-14; see also [9]).

Once we have the Cholesky factorization of the covariance, we minimize the integrand of (76) using standard convex optimization routines. In this way, we can accurately estimate the capacity for arbitrary data manifolds by following the pseudocode in Algorithms (1) and (2).

4 Gaussian Point Cloud Simulation

For these simulations, we used $P = 80$ point cloud manifolds, each made up of the convex hulls of 40 random vectors in $\mathbb{R}^{3800}$, and we averaged results over 5 runs. These vectors were the sum of two Gaussian vectors: a manifold centroid u^μ_0, and a sample-specific vector, x^μ_j. Each manifold was then defined as $M^\mu = \text{conv}\{u^\mu_0 + x^\mu_j : j \in \mathbb{N}^{40}_1\}$. The sample vectors and centroids were each drawn from zero-mean multivariate Gaussian distributions with block covariance matrices with the strength of the off-diagonal terms of both matrices being uniformly scaled from trial to trial (see Supplementary Fig. 1 above for an example). That is, given an intensity γ, we enforced $\overline{\langle u^\mu_0, u^\nu_0 \rangle} = \gamma C^{\mu,\nu}_{cent}$ for $\mu \neq \nu$, while each of the sample vectors satisfied $\overline{\langle x^\mu_j, x^\nu_i \rangle} = \gamma C^{\mu,\nu}_{samp}$ for $\mu \neq \nu$. The average norms, $||u^\mu_0||^2$ and $||x^\mu_i||^2$ (i.e., the diagonal elements of C) were respectively fixed to 25 and 1.

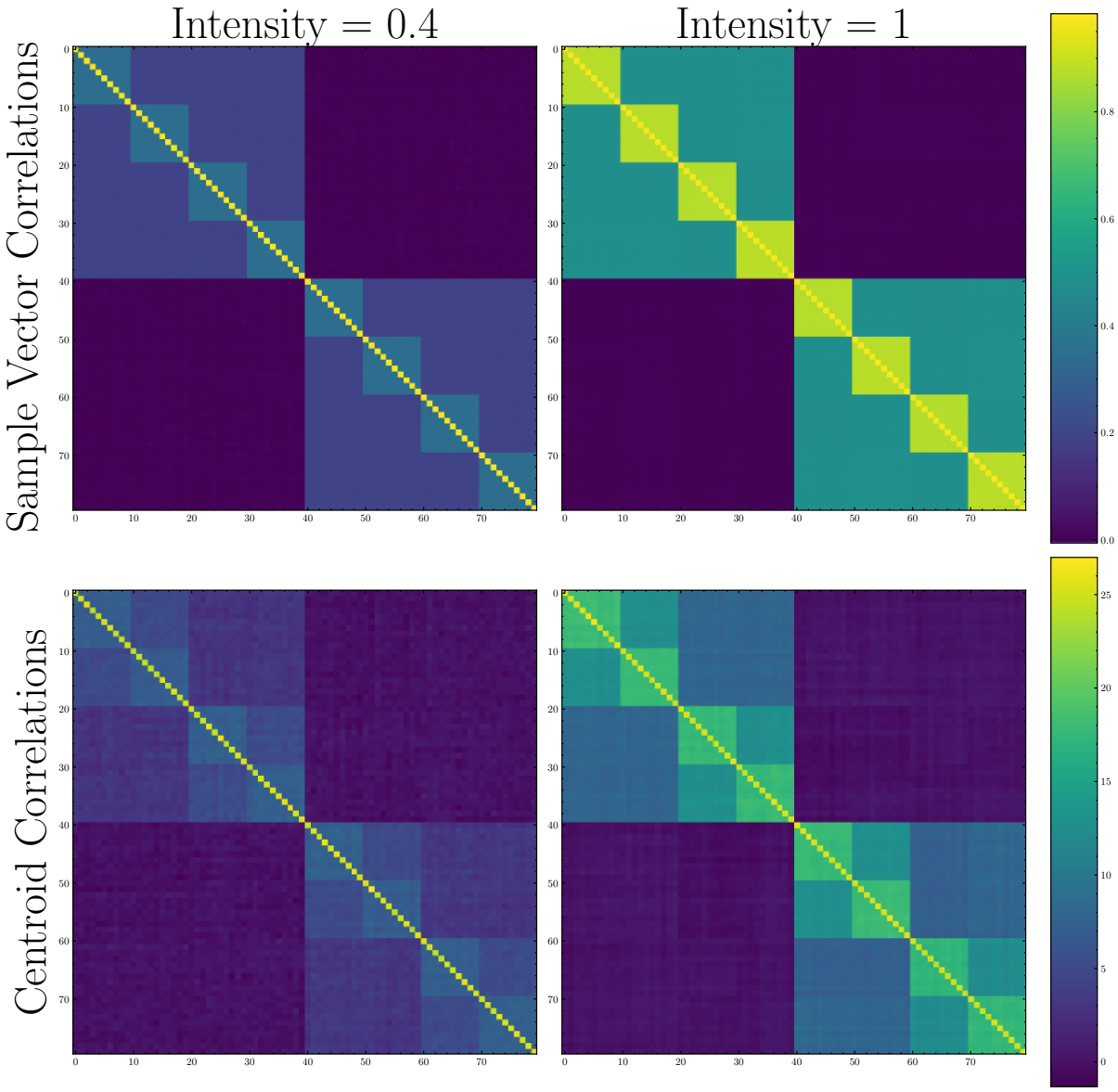

Figure 1: Example of the block covariance matrices for the sample-specific vectors (top row) and centroids (bottom row) at low intensity (left column) and high intensity (right column).

5 ResNet50 Analyses

We generated object manifolds from the ResNet 50 architecture trained with SimCLR [10, 11] using a similar procedure as in [3]. On each of the five experimental runs, we first randomly selected $P = 70$ ImageNet classes and chose 45 samples per class. We then extracted the activations from a sub-sample of 13 out of the 49 ReLU layers in the network, as well as the final average pooling operation. As described in [11], these ReLU layers are the result of applying the rectified linear non-linearity $\mathrm{ReLU}(x) = \max\{0, x\}$ elementwise to the outputs of convolutional layers. The layer width of the final average pooling layer was 2,048, while the layer widths of the ReLU layers ranged from 25,088 to 802,816. Given the size of these layers we projected the activations of the raw input and the ReLU layers onto $N = 8,000$ vectors sampled randomly from the unit sphere in order to conserve memory as in [3]. We then applied the low rank approximation from [3], the simulation capacity algorithm from [12], and our α_{cor} estimator to the projected manifolds. Note that we perform the random projection before performing any of the steps described in Supplementary Section 3. Code reproducing all analyses may be found in [9].

References

[1] E Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, January 1988.

[2] SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Classification and Geometry of General Perceptual Manifolds. *Physical Review X*, 8(3):031003, July 2018.

[3] Uri Cohen, SueYeon Chung, Daniel D. Lee, and Haim Sompolinsky. Separability and geometry of object manifolds in deep neural networks. *Nature Communications*, 11(1):746, February 2020. Number: 1 Publisher: Nature Publishing Group.

[4] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[5] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[6] M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, 1986. _eprint: https://www.worldscientific.com/doi/pdf/10.1142/0271.

[7] William W Hager. Updating the inverse of a matrix. *SIAM review*, 31(2):221–239, 1989.

[8] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[9] Albert Wakhloo, Tamara Sussman, and SueYeon Chung. Capacity for correlated manifolds code. `https://zenodo.org/record/7844169#.ZD9Gwy-B22s`, 2023. 10.5281/zenodo.7844169.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020. tex.organization: PMLR.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] SueYeon Chung, Uri Cohen, Haim Sompolinsky, and Daniel D. Lee. Learning data manifolds with a cutting plane method. *Neural Computation*, 30(10):2593–2615, October 2018.