Editors' Suggestion   Featured in Physics

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, 162 Fifth Avenue, New York, New York 10010, USA*
[2] *Department of Child and Adolescent Psychiatry, New York State Psychiatric Institute,*
1051 Riverside Drive, New York, New York 10032, USA
[3] *Columbia University Irving Medical College, 630 West 168th Street, New York, New York 10032, USA*
[4] *Center for Neural Science, New York University, 4 Washington Place, New York, New York 10003, USA*

(Received 29 November 2022; revised 3 March 2023; accepted 21 April 2023; published 12 July 2023)

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.

DOI: 10.1103/PhysRevLett.131.027301

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 Refs. [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 Ref. [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 [18–21]. 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 [22–25]. These considerations call for a theoretical characterization of the relationship between network performance, representational geometry, and the correlation structure of network representations.

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 Ref. [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

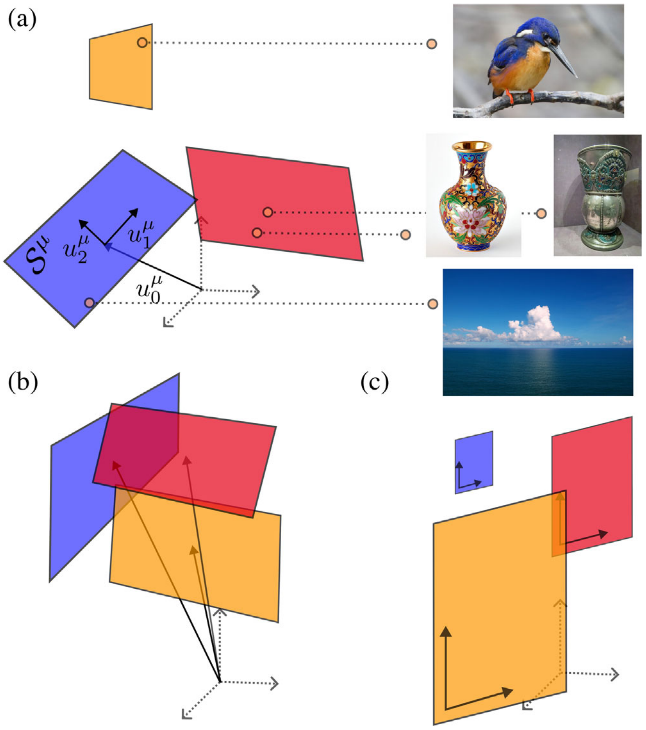

FIG. 1. (a) Neural manifolds arising from different instances of $P = 3$ object classes (bird, vase, and cloud [26–29]), with $N = 3$ neurons. We parametrize the manifolds in terms of a centroid u_0^μ, axes $u_{i>0}^\mu$, and shape vectors $\mathcal{S}^\mu$, determining which linear combinations of the 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.

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 $\min_{x\in M^\mu} y^\mu \langle w, x\rangle \geq \kappa$ 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 Refs. [2,15,30–34], we study this problem by calculating the average log-volume of the space of solutions in the thermodynamic limit:

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

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, $\mathcal{S}^\mu$:

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

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: $p(u) \propto \exp[-(N/2) \sum_{\mu,\nu,i,j,l} (C^{-1})_{\nu,j}^{\mu,i} u_{i,l}^\mu u_{j,l}^\nu]$.

We calculate the capacity for correlated manifolds using the replica method [35,36]; the details can be found in the

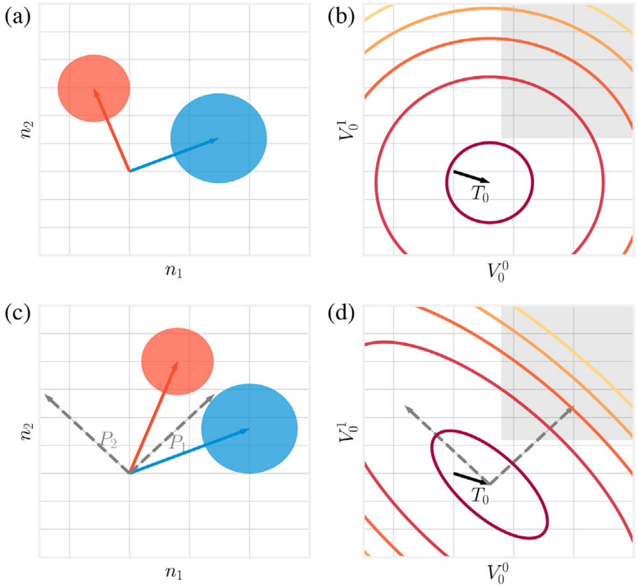

FIG. 2. The effect of correlations on the optimization landscape for V_0. *First column:* Two manifolds with (a) uncorrelated and (c) correlated centroids arising from the activations of two neurons, n_1 and n_2. *Second column:* Level curves for $\|V-T\|^2_{y,C}$, given fixed y and $V_{i>0}$ for the (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 $y^\mu y^\nu C^{\mu,0}_{\nu,0}$.

Supplemental Material [37]. We find that the capacity at a margin κ, denoted by $\alpha_{cor}(\kappa)$, is

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

where $D_{y,C}T$ is the zero-mean Gaussian measure with covariance tensor $y^\mu y^\nu C^{\mu,i}_{\nu,j}$, and the overline denotes the remaining average with respect to the labels y^μ. Note too that we have defined the Mahalanobis norm: $\|X\|^2_{y,C} \equiv \sum_{\mu,\nu,i,j} y^\mu y^\nu (C^{-1})^{\mu,i}_{\nu,j} X^\mu_i X^\nu_j$, which effectively stretches 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 Refs. [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 Ref. [16] when applied to Gaussian point cloud manifolds (Fig. 3). Therefore, the correlated capacity estimator, α_{cor}, whose numerical implementation we describe in the Supplemental Material [37], should be used whenever working with manifolds with strong correlations (see Ref. [39]).

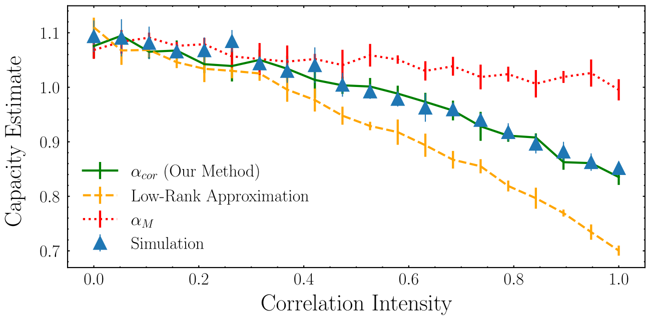

FIG. 3. Comparison of three different capacity estimators, including the low-rank approximation of Ref. [16], to the numerically estimated ground truth simulation capacity (blue triangles) described in Ref. [40]. The correlation intensity denotes the magnitude of the off-diagonal correlations—see the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

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

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

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 λ, $\psi \to 1$, the axes and centroids will be fully correlated with one another, while λ, $\psi \to 0$ 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 the Supplemental Material [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: $[\alpha_{cor}(\kappa)]^{-1} = -(2/P)\lim_{\beta\to\infty}(\partial/\partial\beta)\overline{\log Z}$, 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 the Supplemental Material for details [37].

Under these assumptions, the capacity is given by

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

where q is the scaled squared norm of the signed fields of an arbitrary sphere, $q \equiv \overline{\sum_{i>0}(V^\mu_i)^2/(K(1-\lambda))}$, 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

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

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 Refs. [4,34] for related results.

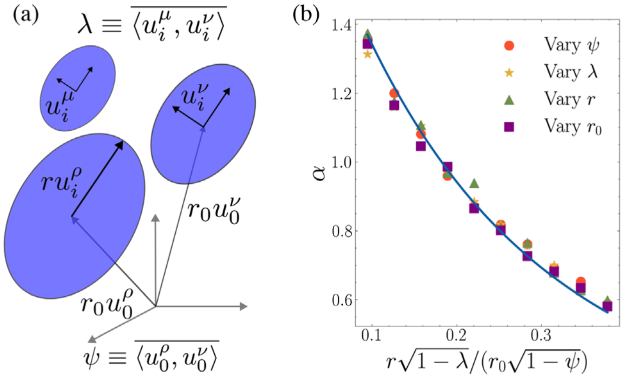

FIG. 4. The capacity for correlated spheres. (a) Visual demonstration of spherical manifolds with low-rank axis and centroid correlations. (b) The zero-margin capacity as a function of only the input ratio $r\sqrt{1-\lambda}/(r_0\sqrt{1-\psi})$. Points represent averages over five random sphere samplings, 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.

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 [22,43,44]. We can see from Fig. 5 that the low rank approximation [16] significantly overestimates the capacity in later layers of the

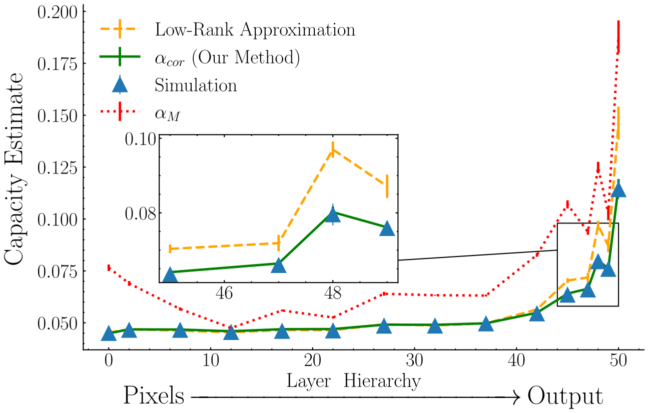

FIG. 5. Comparison of the low-rank approximation (yellow dashed line) [16], α_M (red dotted line) [15], and our α_{cor} calculation (green solid line) to the ground truth simulation capacity (blue triangles) [17] on data manifolds arising from the ResNet50 artificial neural network architecture trained using SimCLR on the ImageNet dataset [22,43].

network. Note that while we can numerically estimate the ground truth simulation capacity here because we use few data points (see the Supplemental Material [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 previous 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 Letter 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 [18,19], 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 preexisting 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.

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, Phys. Rev. Res. **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, Phys. Rev. Lett. **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, Phys. Rev. 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, Phys. Rev. Res. **4**, 023255 (2022).

[6] L. Susman, F. Mastrogiuseppe, N. Brenner, and O. Barak, Quality of internal representation shapes learning performance in feedback neural networks, Phys. Rev. Res. **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* edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché Buc, E. Fox, and R. Garnett (Curran Associates, Inc., Vancouver, 2019), Vol. 32.

[8] D. Dahmen, M. Gilson, and M. Helias, Capacity of the covariance perceptron, J. Phys. A **53**, 354002 (2020).

[9] J. Steinberg and H. Sompolinsky, Associative memory of knowledge structures and sequences, Sci. Rep. **12**, 21808 (2022).

[10] U. Cohen and H. Sompolinsky, Soft-margin classification of object manifolds, Phys. Rev. E **106**, 024126 (2022).

[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, Nat. Neurosci. **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).

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

[14] B. Sorscher, S. Ganguli, and H. Sompolinsky, Neural representational geometry underlies few-shot concept learning, Proc. Natl. Acad. Sci. U.S.A. **119**, e2200800119 (2022).

[15] S. Chung, D. D. Lee, and H. Sompolinsky, Classification and Geometry of General Perceptual Manifolds, Phys. Rev. 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, Nat. Commun. **11**, 746 (2020).

[17] S. Chung, D. D. Lee, and H. Sompolinsky, Linear readout of object manifolds, Phys. Rev. E **93**, 060301 (2016).

[18] S. Panzeri, M. Moroni, H. Safaai, and C. D. Harvey, The structures and functions of correlations in neural population codes, Nat. Rev. Neurosci. **23**, 551 (2022).

[19] J. Zylberberg, A. Pouget, P. E. Latham, and E. Shea-Brown, Robust information propagation through noisy neural circuits, PLoS Comput. Biol. **13**, e1005497 (2017).

[20] A. Morcos, M. Raghu, and S. Bengio, Insights on representational similarity in neural networks with canonical correlation, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Montreal, Quebec, 2018).

[21] 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, https://proceedings.mlr.press/v97/kornblith19a.html.

[22] 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, https://proceedings.mlr.press/v119/chen20j.html.

[23] A. Bardes, J. Ponce, and Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, in *International Conference on Learning Representations* (2022), https://openreview.net/forum?id=xm6YD62D1Ub.

[24] 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, https://proceedings.mlr.press/v139/zbontar21a.html.

[25] 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.

[26] J. J. Harrison, Azure Kingfisher (2011), https://upload.wikimedia.org/wi kipedia/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.

[27] 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.

[28] 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.

[29] 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.

[30] E. Gardner, The space of interactions in neural network models, J. Phys. A **21**, 257 (1988).

[31] R. Rubin, R. Monasson, and H. Sompolinsky, Theory of Spike Timing-Based Neural Classifiers, Phys. Rev. Lett. **105**, 218102 (2010).

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

[33] R. Monasson, Properties of neural networks storing spatially correlated patterns, J. Phys. A **25**, 3701 (1992).

[34] B Lopez, M Schroder, and M Opper, Storage of correlated patterns in a perceptron, J. Phys. A **28**, L447 (1995).

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

[36] M. Mezard and A. Montanari, *Information, Physics, and Computation* (Oxford University Press, New York, 2009).

[37] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.131.027301 for details of the replica calculations and experimental details, which includes Ref. [38].

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

[39] A. Wakhloo, T. Sussman, and S. Chung, *Capacity for correlated manifolds code* (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 Comput. **30**, 2593 (2018).

[41] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, England, 2004).

[42] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A **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, https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

[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, Int. J. Comput. Vision **115**, 211 (2015).

[45] A. M. Saxe, J. L. McClelland, and S. Ganguli, A mathematical theory of semantic development in deep neural networks, Proc. Natl. Acad. Sci. U.S.A. **116**, 11537 (2019).

[46] W. J. Johnston and S. Fusi, Abstract representations emerge naturally in neural networks trained to perform multiple tasks, Nat. Commun. **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), https://openreview.net/forum?id=Sy2fzU9gl.

[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.