Classification and Geometry of General Perceptual Manifolds

SueYeon Chung,[1,2,6] Daniel D. Lee,[2,3] and Haim Sompolinsky[2,4,5]

[1]*Program in Applied Physics, School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*
[3]*School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA*
[4]*Racah Institute of Physics, Hebrew University, Jerusalem 91904, Israel*
[5]*Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 91904, Israel*
[6]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*

Perceptual manifolds arise when a neural population responds to an ensemble of sensory signals associated with different physical features (e.g., orientation, pose, scale, location, and intensity) of the same perceptual object. Object recognition and discrimination require classifying the manifolds in a manner that is insensitive to variability within a manifold. How neuronal systems give rise to invariant object classification and recognition is a fundamental problem in brain theory as well as in machine learning. Here, we study the ability of a readout network to classify objects from their perceptual manifold representations. We develop a statistical mechanical theory for the linear classification of manifolds with arbitrary geometry, revealing a remarkable relation to the mathematics of conic decomposition. We show how special anchor points on the manifolds can be used to define novel geometrical measures of radius and dimension, which can explain the classification capacity for manifolds of various geometries. The general theory is demonstrated on a number of representative manifolds, including ℓ_2 ellipsoids prototypical of strictly convex manifolds, ℓ_1 balls representing polytopes with finite samples, and ring manifolds exhibiting nonconvex continuous structures that arise from modulating a continuous degree of freedom. The effects of label sparsity on the classification capacity of general manifolds are elucidated, displaying a universal scaling relation between label sparsity and the manifold radius. Theoretical predictions are corroborated by numerical simulations using recently developed algorithms to compute maximum margin solutions for manifold dichotomies. Our theory and its extensions provide a powerful and rich framework for applying statistical mechanics of linear classification to data arising from perceptual neuronal responses as well as to artificial deep networks trained for object recognition tasks.

Subject Areas: Biological Physics, Complex Systems, Statistical Physics

I. INTRODUCTION

One fundamental cognitive task performed by animals and humans is the invariant perception of objects, requiring the nervous system to discriminate between different objects despite substantial variability in each object's physical features. For example, in vision, the mammalian brain is able to recognize objects despite variations in their orientation, position, pose, lighting, and background. Such impressive robustness to physical changes is not limited to

vision; other examples include speech processing, which requires the detection of phonemes despite variability in the acoustic signals associated with individual phonemes, and the discrimination of odors in the presence of variability in odor concentrations. Sensory systems are organized as hierarchies, consisting of multiple layers, transforming sensory signals into a sequence of distinct neural representations. Studies of high-level sensory systems, e.g., the inferotemporal cortex (IT) in vision [1], auditory cortex in audition [2], and piriform cortex in olfaction [3], reveal that even the late sensory stages exhibit significant sensitivity of neuronal responses to physical variables. This suggests that sensory hierarchies generate representations of objects that, although not entirely invariant to changes in physical features, are still readily decoded in an invariant manner by a downstream system. This hypothesis is formalized
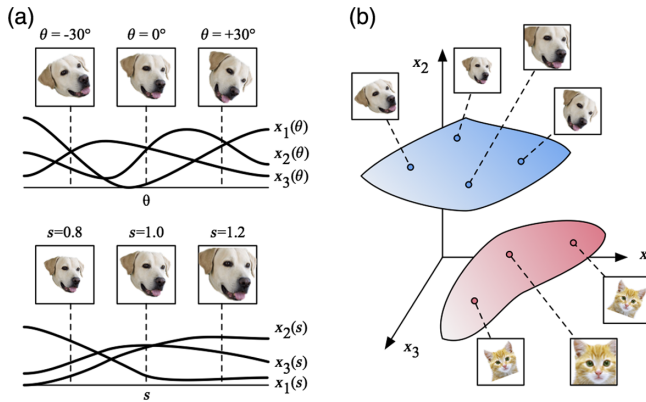
FIG. 1. Perceptual manifolds in neural state space.(a) Firing rates of neurons responding to images of a dog shown at various orientations θ and scales s. The response to a particular orientation and scale can be characterized by an N-dimensional population response. (b) The population responses to the images of the dog form a continuous manifold representing the complete set of invariances in the $\mathbb{R}^N$ neural activity space. Other object images, such as those corresponding to a cat in various poses, are represented by other manifolds in this vector space.

by the notion of the untangling of perceptual manifolds [4–6]. This viewpoint underlies a number of studies of object recognition in deep neural networks for artificial intelligence [7–10].

To conceptualize perceptual manifolds, consider a set of N neurons responding to a specific sensory signal associated with an object, as shown in Fig. 1. The neural population response to that stimulus is a vector in $\mathbb{R}^N$. Changes in the physical parameters of the input stimulus that do not change the object identity modulate the neural state vector. The set of all state vectors corresponding to responses to all possible stimuli associated with the same object can be viewed as a manifold in the neural state space. In this geometrical perspective, object recognition is equivalent to the task of discriminating manifolds of different objects from each other. Presumably, as signals propagate from one processing stage to the next in the sensory hierarchy, the geometry of the manifolds is reformatted so that they become "untangled"; namely, they are more easily separated by a biologically plausible decoder [1]. In this paper, we model the decoder as a simple single layer network (the perceptron) and ask how the geometrical properties of the perceptual manifolds influence their ability to be separated by a linear classifier.

Linear separability has previously been studied in the context of the classification of points by a perceptron, using combinatorics [11] and statistical mechanics [12,13]. Gardner's statistical mechanics theory is extremely important as it provides accurate estimates of the perceptron capacity beyond function counting by incorporating robustness measures. The robustness of a linear classifier is quantified by the margin, which measures the distance between the separating hyperplane and the closest point. Maximizing the margin of a classifier is a critical

objective in machine learning, providing support vector machines (SVM) with their good generalization performance guarantees [14].

The above theories focus on separating a finite set of points with no underlying geometrical structure and are not applicable to the problem of manifold classification, which deals with separating an infinite number of points geometrically organized as manifolds. This paper addresses the important question of how to quantify the capacity of the perceptron for dichotomies of input patterns described by manifolds. In an earlier paper, we have presented the analysis for classification of manifolds of extremely simple geometry, namely, balls [15]. However, the previous results have limited applicability as the neural manifolds arising from realistic physical variations of objects can exhibit much more complicated geometries. Can statistical mechanics deal with the classification of manifolds with complex geometry, and what specific geometric properties determine the separability of manifolds?

In this paper, we develop a theory of the linear separability of general, finite-dimensional manifolds. The summary of our key results is as follows:

  (i) We begin by introducing a mathematical model of general manifolds for binary classification (Sec. II). This formalism allows us to generate generic bounds on the manifold separability capacity from the limits of small manifold sizes (classification of isolated points) to that of large sizes (classification of entire affine subspaces). These bounds highlight the fact that for large ambient dimension N, the maximal number P of separable finite-dimensional manifolds is proportional to N, even though each consists of an infinite number of points, setting the stage for a statistical mechanical evaluation of the maximal $\alpha = (P/N)$.

 (ii) Using replica theory, we derive mean-field equations for the capacity of linear separation of finite-dimensional manifolds (Sec. III) and for the statistical properties of the optimal separating weight vector. The mean-field solution is given in the form of self-consistent Karush-Kuhn-Tucker (KKT) conditions involving the manifold anchor point. The anchor point is a representative support vector for the manifold. The position of the anchor point on a manifold changes as the orientations of the other manifolds are varied, and the ensuing statistics of the distribution of anchor points plays a key role in our theory. The optimal separating plane intersects a fraction of the manifolds (the supporting manifolds). Our theory categorizes the dimension of the span of the intersecting sets (e.g., points, edges, faces, or full manifolds) in relation to the position of the anchor points in the manifolds' convex hulls.

(iii) The mean-field theory motivates a new definition of manifold geometry, which is based on the measure induced by the statistics of the anchor points.

In particular, we define the manifold anchor radius and dimension, R_M and D_M, respectively. These quantities are relevant since the capacity of general manifolds can be well approximated by the capacity of ℓ_2 balls with radii R_M and dimensions D_M. Interestingly, we show that, in the limit of small manifolds, the anchor point statistics are dominated by points on the boundary of the manifolds which have minimal overlap with Gaussian random vectors. The resultant Gaussian radius R_g and dimension D_g are related to the well-known *Gaussian mean width* of convex bodies (Sec. IV). Beyond understanding fundamental limits for classification capacity, these geometric measures offer new quantitative tools for assessing how perceptual manifolds are reformatted in brain and artificial systems.

(iv) We apply the general theory to three examples, representing distinct prototypical manifold classes. One class consists of manifolds with strictly smooth convex hulls, which do not contain facets and are exemplified by ℓ_2 ellipsoids. Another class is that of convex polytopes, which arise when the manifolds consist of a finite number of data points, and are exemplified by ℓ_1 ellipsoids. Finally, ring manifolds represent an intermediate class: smooth but non-convex manifolds. Ring manifolds are continuous nonlinear functions of a single intrinsic variable, such as object orientation angle. The differences between these manifold types show up most clearly in the distinct patterns of the support dimensions. However, as we show, they share common trends. As the size of the manifold increases, the capacity and geometrical measures vary smoothly, exhibiting a smooth crossover from a small radius and dimension with high capacity to a large radius and dimension with low capacity. This crossover occurs as $R_g \propto (1/\sqrt{D_g})$. Importantly, for many realistic cases, when the size is smaller than the crossover value, the manifold dimensionality is substantially smaller than that computed from naive second-order statistics, highlighting the saliency and significance of our measures for the anchor geometry.

(v) Finally, we treat the important case of the classification of manifolds with imbalanced (sparse) labels, which commonly arise in problems of object recognition. It is well known that, in highly sparse labels, the classification capacity of random points increases dramatically as $(f|\log f|)^{-1}$, where $f \ll 1$ is the fraction of the minority labels. Our analysis of sparsely labeled manifolds highlights the interplay between manifold size and sparsity. In particular, it shows that sparsity enhances the capacity only when $fR_g^2 \ll 1$, where R_g is the (Gaussian) manifold radius. Notably, for a large regime of parameters, sparsely labeled manifolds can approximately be described by a universal capacity function

equivalent to sparsely labeled ℓ_2 balls with radii R_g and dimensions D_g, as demonstrated by our numerical evaluations (Sec. VI). Conversely, when $fR_g^2 \gg 1$, the capacity is low and close to $(1/D)$, where D is the dimensionality of the manifold affine subspace, even for extremely small f.

(vi) Our theory provides, for the first time, quantitative and qualitative predictions for the perceptron classification of realistic data structures. However, application to real data may require further extensions of the theory and are discussed in Sec. VII. Together, the theory makes an important contribution to the development of statistical mechanical theories of neural information processing in realistic conditions.

II. MODEL OF MANIFOLDS

Manifolds in affine subspaces:—We model a set of P perceptual manifolds corresponding to P perceptual objects. Each manifold M^μ for $\mu = 1, \ldots, P$ consists of a compact subset of an affine subspace of $\mathbb{R}^N$ with *affine dimension D* with $D < N$. A point on the manifold $\mathbf{x}^\mu \in M^\mu$ can be parametrized as

$$\mathbf{x}^\mu(\vec{S}) = \sum_{i=1}^{D+1} S_i \mathbf{u}_i^\mu, \qquad (1)$$

where $\mathbf{u}_i^\mu$ are a set of orthonormal bases of the $(D+1)$-dimensional linear subspace containing M^μ. The $D+1$ components S_i represent the coordinates of the manifold point within this subspace and are constrained to be in the set $\vec{S} \in \mathcal{S}$. The bold notation for $\mathbf{x}^\mu$ and $\mathbf{u}_i^\mu$ indicates that they are vectors in $\mathbb{R}^N$, whereas the arrow notation for $\vec{S}$ indicates that it is a vector in $\mathbb{R}^{D+1}$. The set $\mathcal{S}$ defines the shape of the manifolds and encapsulates the affine constraint. For simplicity, we will first assume that the manifolds have the same geometry so that the coordinate set $\mathcal{S}$ is the same for all the manifolds; extensions that consider heterogeneous geometries are provided in Sec. VI A.

We study the separability of P manifolds into two classes, denoted by binary labels $y^\mu = \pm 1$, by a linear hyperplane passing through the origin. A hyperplane is described by a weight vector $\mathbf{w} \in \mathbb{R}^N$, normalized so $\|\mathbf{w}\|^2 = N$, and the hyperplane correctly separates the manifolds with a margin $\kappa \geq 0$ if it satisfies

$$y^\mu \mathbf{w} \cdot \mathbf{x}^\mu \geq \kappa \qquad (2)$$

for all μ and $\mathbf{x}^\mu \in M^\mu$. Since linear separability is a convex problem, separating the manifolds is equivalent to separating the convex hulls, $\text{conv}(M^\mu) = \{\mathbf{x}^\mu(\vec{S}) | \vec{S} \in \text{conv}(\mathcal{S})\}$, where

$$\text{conv}(\mathcal{S}) = \left\{ \sum_{i=1}^{D+1} \alpha_i \vec{S}_i | \vec{S}_i \in \mathcal{S}, \alpha_i \geq 0, \sum_{i=1}^{D+1} \alpha_i = 1 \right\}. \qquad (3)$$
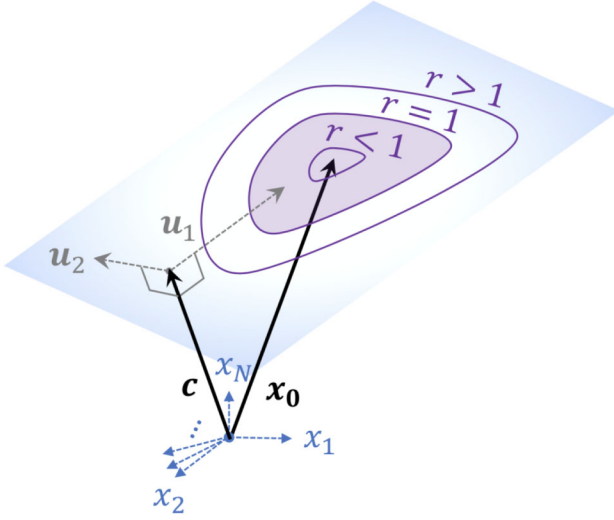
FIG. 2. Model of manifolds in affine subspaces. $D = 2$ manifold embedded in $\mathbb{R}^N$. $\mathbf{c}$ is the orthogonal translation vector for the affine space, and $\mathbf{x_0}$ is the center of the manifold. As the scale r is varied, the manifold shrinks to the point $\mathbf{x_0}$ or expands to fill the entire affine space.

The position of an affine subspace relative to the origin can be defined via the translation vector that is closest to the origin. This orthogonal translation vector $\mathbf{c}^\mu$ is perpendicular to all the affine displacement vectors in M^μ, so that all the points in the affine subspace have equal projections on $\mathbf{c}^\mu$, i.e., $\vec{\mathbf{x}}^\mu \cdot \mathbf{c}^\mu = \|\mathbf{c}^\mu\|^2$ for all $\mathbf{x}^\mu \in M^\mu$ (Fig. 2). We will further assume for simplicity that the P norms $\|\mathbf{c}^\mu\|$ are all the same and normalized to 1.

To investigate the separability properties of manifolds, it is helpful to consider scaling a manifold M^μ by an overall scale factor r without changing its shape. We define the scaling relative to a center $\vec{S}^0 \in \mathcal{S}$ by a scalar $r > 0$, by

$$rM^\mu = \left\{ \sum_{i=1}^{D+1} [S_i^0 + r(S_i - S_i^0)] \mathbf{u}_i^\mu | \vec{S} \in \mathcal{S} \right\}. \quad (4)$$

When $r \to 0$, the manifold rM^μ converges to a point: $\mathbf{x}_0^\mu = \sum_i S_i^0 \mathbf{u}_i^\mu$. On the other hand, when $r \to \infty$, the manifold rM^μ spans the entire affine subspace. If the manifold is symmetric (such as for an ellipsoid), there is a natural choice for a center. We will later provide an appropriate definition for the center point for general, asymmetric manifolds. In general, the translation vector $\vec{c}$ and center $\vec{S}^0$ need not coincide, as shown in Fig. 2. However, we will also discuss later the special case of centered manifolds, in which the translation vector and center do coincide.

Bounds on linear separability of manifolds:—For dichotomies of P input points in $\mathbb{R}^N$ at zero margin, $\kappa = 0$, the number of dichotomies that can be separated by a linear hyperplane through the origin is given by [11]

$$C_0(P, N) = 2 \sum_{k=0}^{N-1} C_k^{P-1} \leq 2^P, \quad (5)$$

where $C_k^n = [n!/k!(n-k)!]$ is the binomial coefficient for $n \geq k$, and zero otherwise. This result holds for P input vectors that obey the mild condition that the vectors are in general position, namely, that all subsets of input vectors of size $p \leq N$ are linearly independent.

For large P and N, the probability $(1/2^P)C_0(P, N)$ of a dichotomy being linearly separable depends only upon the ratio (P/N) and exhibits a sharp transition at the critical value of $\alpha_0 = 2$. We are not aware of a comprehensive extension of Cover's counting theorem for general manifolds; nevertheless, we can provide lower and upper bounds on the number of linearly realizable dichotomies by considering the limit of $r \to 0$ and $r \to \infty$ under the following general conditions.

First, in the limit of $r \to 0$, the linear separability of P manifolds becomes equivalent to the separability of the P centers. This leads to the requirement that the centers of the manifolds $\mathbf{x}_0^\mu$ are in a general position in $\mathbb{R}^N$. Second, we consider the conditions under which the manifolds are linearly separable when $r \to \infty$, so that the manifolds span complete affine subspaces. For a weight vector $\mathbf{w}$ to consistently assign the same label to all points on an affine subspace, it must be orthogonal to all the displacement vectors in the affine subspace. Hence, to realize a dichotomy of P manifolds when $r \to \infty$, the weight vector $\mathbf{w}$ must lie in a null space of dimension $N - D_{\text{tot}}$, where D_{tot} is the rank of the union of affine displacement vectors. When the basis vectors $\mathbf{u}_i^\mu$ are in general position, then $D_{\text{tot}} = \min(DP, N)$. Then, for the affine subspaces to be separable, $PD < N$ is required, and the projections of the P orthogonal translation vectors also need to be separable in the $(N - D_{\text{tot}})$-dimensional null space. Under these general conditions, the number of dichotomies for D-dimensional affine subspaces that can be linearly separated, $C_D(P, N)$, can be related to the number of dichotomies for a finite set of points via

$$C_D(P, N) = C_0(P, N - PD). \quad (6)$$

From this relationship, we conclude that the ability to linearly separate D-dimensional affine subspaces exhibits a transition from always being separable to never being separable at the critical ratio $(P/N) = (2/1 + 2D)$ for large P and N [see Supplemental Materials (SM) [16], Sec. S1].

For general D-dimensional manifolds with finite size, the number of dichotomies that are linearly separable will be lower bounded by $C_D(P, N)$ and upper bounded by $C_0(P, N)$. We introduce the notation $\alpha_M(\kappa)$ to denote the maximal load (P/N) such that randomly labeled manifolds are linearly separable with a margin κ, with high probability. Therefore, from the above considerations, it follows that the critical load at zero margin, $\alpha_M(\kappa = 0)$, is bounded by

$$\frac{1}{2} \leq \alpha_M^{-1}(\kappa = 0) \leq \frac{1}{2} + D. \quad (7)$$

These bounds highlight the fact that, in the large N limit, the maximal number of separable finite-dimensional manifolds is proportional to N, even though each consists of an infinite number of points. This sets the stage for a statistical mechanical evaluation of the maximal $\alpha = (P/N)$, where P is the number of manifolds, and is described in the following section.

III. STATISTICAL MECHANICAL THEORY

In order to make theoretical progress beyond the bounds above, we need to make additional statistical assumptions about the manifold spaces and labels. Specifically, we will assume that the individual components of $\mathbf{u}_i^\mu$ are drawn independently and from identical Gaussian distributions with zero mean and variance $(1/N)$, and that the binary labels $y^\mu = \pm 1$ are randomly assigned to each manifold with equal probabilities. We will study the thermodynamic limit where $N, P \to \infty$, but with a finite load $\alpha = (P/N)$. In addition, the manifold geometries as specified by the set $\mathcal{S}$ in $\mathbb{R}^{D+1}$ and, in particular, their affine dimension D are held fixed in the thermodynamic limit. Under these assumptions, the bounds in Eq. (7) can be extended to the linear separability of general manifolds with finite margin κ, and characterized by the reciprocal of the critical load ratio $\alpha_M^{-1}(\kappa)$,

$$\alpha_0^{-1}(\kappa) \leq \alpha_M^{-1}(\kappa) \leq \alpha_0^{-1}(\kappa) + D, \qquad (8)$$

where $\alpha_0(\kappa)$ is the maximum load for separation of random independent and identically distributed (i.i.d.) points, with a margin κ given by the Gardner theory [12],

$$\alpha_0^{-1}(\kappa) = \int_{-\infty}^{\kappa} Dt(t-\kappa)^2, \qquad (9)$$

with Gaussian measure $Dt = (1/\sqrt{2\pi})e^{-(t^2/2)}$. For many interesting cases, the affine dimension D is large, and the gap in Eq. (8) is overly loose. Hence, it is important to derive an estimate of the capacity for manifolds with finite sizes and evaluate the dependence of the capacity and the nature of the solution on the geometrical properties of the manifolds, as shown below.

A. Mean-field theory of manifold separation capacity

Following Gardner's framework [12,13], we compute the statistical average of $\log Z$, where Z is the volume of the space of the solutions, which in our case can be written as

$$Z = \int d^N \mathbf{w} \delta(\|\mathbf{w}\|^2 - N)\Pi_{\mu, \mathbf{x}^\mu \in M^\mu}\Theta(y^\mu \mathbf{w} \cdot \mathbf{x}^\mu - \kappa). \qquad (10)$$

$\Theta(\cdot)$ is the Heaviside function to enforce the margin constraints in Eq. (2), along with the delta function to ensure $\|\mathbf{w}\|^2 = N$. In the following, we focus on the

properties of the maximum margin solution, namely, the solution for the largest load α_M for a fixed margin κ, or equivalently, the solution when the margin κ is maximized for a given α_M.

As shown in Appendix A, we prove that the general form of the inverse capacity, exact in the thermodynamic limit, is

$$\alpha_M^{-1}(\kappa) = \langle F(\vec{T})\rangle_{\vec{T}}, \qquad (11)$$

where $F(\vec{T}) = \min_{\vec{V}}\{\|\vec{V} - \vec{T}\|^2 | \vec{V} \cdot \vec{S} - \kappa \geq 0, \ \forall \ \vec{S} \in \mathcal{S}\}$ and $\langle \ldots \rangle_{\vec{T}}$ is an average over random $(D + 1)$-dimensional vectors $\vec{T}$, whose components are i.i.d. normally distributed $T_i \sim \mathcal{N}(0, 1)$. The components of the vector $\vec{V}$ represent the signed fields induced by the solution vector $\mathbf{w}$ on the $D + 1$ basis vectors of the manifold. The Gaussian vector $\vec{T}$ represents the part of the variability in $\vec{V}$ due to quenched variability in the manifold's basis vectors and the labels, as will be explained in detail below.

The inequality constraints in F can be written equivalently, as a constraint on the point on the manifold with minimal projection on $\vec{V}$. We therefore consider the concave support function of $\mathcal{S}$, $g_{\mathcal{S}}(\vec{V}) = \min_{\vec{S}}\{\vec{V} \cdot \vec{S} | \vec{S} \in \mathcal{S}\}$, which can be used to write the constraint for $F(\vec{T})$ as

$$F(\vec{T}) = \min_{\vec{V}}\{\|\vec{V} - \vec{T}\|^2 | g_{\mathcal{S}}(\vec{V}) - \kappa \geq 0\}. \qquad (12)$$

Note that this definition of $g_{\mathcal{S}}(\vec{V})$ is easily mapped to the conventional convex support function defined via the *max* operation [17].

KKT conditions:—To gain insight into the nature of the maximum margin solution, it is useful to consider the KKT conditions of the convex optimization in Eq. (12) [17]. For each $\vec{T}$, the KKT conditions that characterize the unique solution of $\vec{V}$ for $F(\vec{T})$ is given by

$$\vec{V} = \vec{T} + \lambda \tilde{S}(\vec{T}), \qquad (13)$$

where

$$\lambda \geq 0$$
$$g_{\mathcal{S}}(\vec{V}) - \kappa \geq 0$$
$$\lambda[g_{\mathcal{S}}(\vec{V}) - \kappa] = 0. \qquad (14)$$

The vector $\tilde{S}(\vec{T})$ is a *subgradient* of the support function at $\vec{V}$, $\tilde{S}(\vec{T}) \in \partial g_S(\vec{V})$ [17]; i.e., it is a point on the convex hull of S with minimal overlap with $\vec{V}$. When the support function is differentiable, the subgradient $\partial g_S(\vec{V})$ is unique and is equivalent to the *gradient* of the support function,

$$\tilde{S}(\vec{T}) = \nabla g_{\mathcal{S}}(\vec{V}) = \arg\min_{\vec{S} \in \mathcal{S}} \vec{V} \cdot \vec{S}. \qquad (15)$$

Since the support function is positively homogeneous, $g_{\mathcal{S}}(\gamma \vec{V}) = \gamma g_{\mathcal{S}}(\vec{V})$ for all $\gamma > 0$; thus, $\nabla g_{\mathcal{S}}(\vec{V})$ depends only on the unit vector $\hat{V}$. For values of $\vec{V}$ such that $g_{\mathcal{S}}(\vec{V})$ is not differentiable, the subgradient is not unique, but $\tilde{S}(\vec{T})$ is defined uniquely as the particular subgradient that obeys the KKT conditions, Eqs. (13)–(14). In the latter case, $\tilde{S}(\vec{T}) \in \text{conv}(\mathcal{S})$ may not reside in S itself.

From Eq. (13), we see that the capacity can be written in terms of $\tilde{S}(\vec{T})$ as

$$F(\vec{T}) = \|\lambda \tilde{S}(\vec{T})\|^2. \qquad (16)$$

The scale factor λ is either zero or positive, corresponding to whether $g_{\mathcal{S}}(\vec{V}) - \kappa$ is positive or zero. If $g_{\mathcal{S}}(\vec{V}) - \kappa$ is positive, then $\lambda = 0$, meaning that $\vec{V} = \vec{T}$ and $\vec{T} \cdot \tilde{S}(\vec{T}) - \kappa > 0$. If $\vec{T} \cdot \tilde{S}(\vec{T}) - \kappa < 0$, then $\lambda > 0$ and $\vec{V} \neq \vec{T}$. In this case, multiplying Eq. (13) with $\tilde{S}(\vec{T})$ yields $\lambda \|\tilde{S}(\vec{T})\|^2 = -\vec{T} \cdot \tilde{S}(\vec{T}) + \kappa$. Thus, λ obeys the self-consistent equation,

$$\lambda = \frac{[-\vec{T} \cdot \tilde{S}(\vec{T}) + \kappa]_+}{\|\tilde{S}(\vec{T})\|^2}, \qquad (17)$$

where the function $[x]_+ = \max(x, 0)$.

B. Mean-field interpretation of the KKT relations

The KKT relations have a nice interpretation within the framework of the mean-field theory. The maximum margin solution vector can always be written as a linear combination of a set of support vectors. Although there are infinite numbers of input points in each manifold, the solution vector can be decomposed into P vectors, one per manifold,

$$\mathbf{w} = \sum_{\mu=1}^{P} \lambda_\mu y^\mu \tilde{\mathbf{x}}^\mu, \qquad \lambda_\mu \geq 0, \qquad (18)$$

where $\tilde{\mathbf{x}}^\mu \in \text{conv}(M^\mu)$ is a vector in the convex hull of the μth manifold. In the large N limit, these vectors are uncorrelated with each other. Hence, squaring this equation, ignoring correlations between the different contributions (Using the "cavity" method, one can show that the λ's appearing in the mean field KKT equations and the λ's appearing in Eq. (18) differ by an overall global scaling factor which accounts for the presence of correlations between the individual terms in Eq. (18), which we have neglected here for brevity; see [18,19]), yields $\|\mathbf{w}\|^2 = N = \sum_{\mu=1}^{P} \lambda_\mu^2 \|\tilde{S}^\mu\|^2$, where $\tilde{S}^\mu$ are the coordinates of $\tilde{\mathbf{x}}^\mu$ in the μth affine subspace of the μth manifold; see Eq. (1). From this equation, it follows that $\alpha^{-1} = (N/P) = \langle \lambda^2 \|\tilde{S}\|^2 \rangle$, which yields the KKT expression for the capacity; see Eqs. (11) and (16).

The KKT relations above are self-consistent equations for the statistics of λ_μ and $\tilde{S}^\mu$. The mean-field theory derives the appropriate statistics from self-consistent equations of the fields on a *single* manifold. To see this, consider projecting the solution vector $\mathbf{w}$ onto the affine subspace of one of the manifolds, say, $\mu = 1$. We define a $(D + 1)$-dimensional vector $\vec{V}^1$ as $V_i^1 = y^1 \mathbf{w} \cdot \mathbf{u}_i^1$, $i = 1, \ldots, D + 1$, which are the signed fields of the solution on the affine basis vectors of the $\mu = 1$ manifold. Then, Eq. (18) reduces to $\vec{V}^1 = \lambda_1 \tilde{S}^1 + \vec{T}$, where $\vec{T}$ represents the contribution from all the other manifolds and since their subspaces are randomly oriented, this contribution is well described as a random Gaussian vector. Finally, self consistency requires that for fixed $\vec{T}$, $\tilde{S}^1$ is such that it has a minimal overlap with $\vec{V}^1$ and represents a point residing on the margin hyperplane; otherwise, it will not contribute to the max margin solution. Thus, Eq. (13) is just the decomposition of the field induced on a specific manifold into the contribution induced by that specific manifold along with the contributions coming from all the other manifolds. The self-consistent Eqs. (14) and (17) relating λ to the Gaussian statistics of $\vec{T}$ then naturally follow from the requirement that $\tilde{S}^1$ represents a support vector.

C. Anchor points and manifold supports

The vectors $\tilde{\mathbf{x}}^\mu$ contributing to the solution, Eq. (18), play a key role in the theory. We will denote them or, equivalently, their affine subspace components $\tilde{S}^\mu$ as the "manifold anchor points." For a particular configuration of manifolds, the manifolds could be replaced by an equivalent set of P anchor points to yield the same maximum margin solution. It is important to stress, however, that an individual anchor point is determined not only by the configuration of its associated manifold, but also by the random orientations of all the other manifolds. For a fixed manifold, the location of its anchor point will vary with the relative configurations of the other manifolds. This variation is captured in mean-field theory by the dependence of the anchor point $\tilde{S}$ on the random Gaussian vector $\vec{T}$.

In particular, the position of the anchor point in the convex hull of its manifold reflects the nature of the relation between that manifold and the margin planes. In general, a fraction of the manifolds will intersect with the margin hyperplanes; i.e., they have nonzero λ. These manifolds are the support manifolds of the system. The nature of this support varies and can be characterized by the dimension k of the span of the intersecting set of $\text{conv}(\mathcal{S})$ with the margin planes. Some support manifolds, which we call "touching" manifolds, intersect with the margin hyperplane only with their anchor point. They have a support dimension $k = 1$, and their anchor point $\tilde{S}$ is on the boundary of S. The other extreme is "fully supporting" manifolds, which completely reside in the margin hyperplane. They are characterized by $k = D + 1$. In this case, $\vec{V}$ is parallel to the translation vector $\vec{c}$ of S. Hence, all the points in S are

support vectors, and all have the same overlap κ with $\vec{V}$. The anchor point, in this case, is the unique point in the interior of conv($\mathcal{S}$) that obeys the self-consistent equation, Eq. (13), namely, that it balances the contribution from the other manifolds to zero out the components of $\vec{V}$ that are orthogonal to $\vec{c}$. In the case of smooth convex hulls (i.e., when S is strongly convex), no other manifold support configurations exist. For other types of manifolds, there are also "partially supporting" manifolds, whose convex hull intersections with the margin hyperplanes consist of k-dimensional faces with $1 < k < D + 1$. The associated anchor points then reside inside the intersecting face. For instance, $k = 2$ implies that $\tilde{S}$ lies on an edge, whereas $k = 3$ implies that $\tilde{S}$ lies on a planar two-face of the convex hull. Determining the dimension of the support structure that arises for various $\vec{T}$ is explained below.

D. Conic decomposition

The KKT conditions can also be interpreted in terms of the conic decomposition of $\vec{T}$, which generalizes the notion of the decomposition of vectors onto linear subspaces and their null spaces via Euclidean projection. The convex cone of the manifold $\mathcal{S}$ is defined as cone($\mathcal{S}$) = $\{\sum_{i=1}^{D+1} \alpha_i \vec{S}_i | \vec{S}_i \in \mathcal{S}, \alpha_i \geq 0\}$; see Fig. 3. The shifted polar cone of $\mathcal{S}$, denoted $\mathcal{S}_\kappa^\circ$, is defined as the convex set of points given by

$$\mathcal{S}_\kappa^\circ = \{\vec{U} \in \mathbb{R}^{D+1} | \vec{U} \cdot \vec{S} + \kappa \leq 0, \quad \forall \ \vec{S} \in \mathcal{S}\} \quad (19)$$

and is illustrated for $\kappa = 0$ and $\kappa > 0$ in Fig. 3. For $\kappa = 0$, Eq. (19) is simply the conventional polar cone of $\mathcal{S}$ [20]. Equation (13) can then be interpreted as the decomposition of $-\vec{T}$ into the sum of two component vectors: One component is $-\vec{V}$, i.e., its Euclidean projection onto $\mathcal{S}_\kappa^\circ$; the other component $\lambda\tilde{S}(\vec{T})$ is located in cone($\mathcal{S}$). When
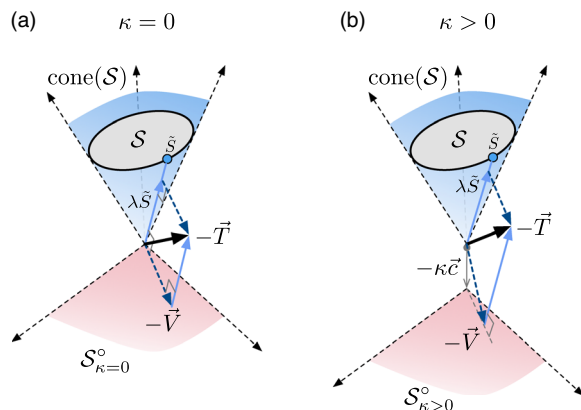

FIG. 3. Conic decomposition of $-\vec{T} = -\vec{V} + \lambda\tilde{S}$ at (a) zero margin $\kappa = 0$ and (b) nonzero margin $\kappa > 0$. Given a random Gaussian $\vec{T}$, $\vec{V}$ is found such that $\|\vec{V} - \vec{T}\|$ is minimized, while $-\vec{V}$ is on the polar cone $\mathcal{S}_\kappa^\circ$. $\lambda\tilde{S}$ is in cone($\mathcal{S}$) and $\tilde{S}$ is the anchor point, a projection on the convex hull of $\mathcal{S}$.

$\kappa = 0$, the Moreau decomposition theorem states that the two components are perpendicular: $\vec{V} \cdot (\lambda\tilde{S}(\vec{T})) = 0$ [21]. For nonzero κ, the two components need not be perpendicular but obey $\vec{V} \cdot (\lambda\tilde{S}(\vec{T})) = \kappa\lambda$.

The position of vector $\vec{T}$ in relation to the cones, cone($\mathcal{S}$) and $\mathcal{S}_\kappa^\circ$, gives rise to qualitatively different expressions for $F(\vec{T})$ and contributions to the solution weight vector and inverse capacity. These correspond to the different support dimensions mentioned above. In particular, when $-\vec{T}$ lies inside $\mathcal{S}_\kappa^\circ$, $\vec{T} = \vec{V}$ and $\lambda = 0$, so the support dimension $k = 0$. On the other hand, when $-\vec{T}$ lies inside cone($\mathcal{S}$), $\vec{V} = \kappa\vec{c}$, and the manifold is fully supporting, $k = D + 1$.

E. Numerical solution of the mean-field equations

The solution of the mean-field equations consists of two stages. First, $\tilde{S}$ is computed for a particular $\vec{T}$, and then the relevant contributions to the inverse capacity are averaged over the Gaussian distribution of $\vec{T}$. For simple geometries such as ℓ_2 ellipsoids, the first step may be solved analytically. However, for more complicated geometries, both steps need to be performed numerically. The first step involves determining $\vec{V}$ and $\tilde{S}$ for a given $\vec{T}$ by solving the quadratic semi-infinite programming problem (QSIP), Eq. (12), over the manifold $\mathcal{S}$, which may contain infinitely many points. A novel "cutting plane" method has been developed to efficiently solve the QSIP problem; see SM [16] (Sec. S4). Expectations are computed by sampling the Gaussian $\vec{T}$ in $D + 1$ dimensions and taking the appropriate averages, similar to procedures for other mean-field methods. The relevant quantities corresponding to the capacity are quite concentrated and converge quickly with relatively few samples.

In the following sections, we will also show how the mean-field theory compares with computer simulations that numerically solve for the maximum margin solution of realizations of P manifolds in $\mathbb{R}^N$, given by Eq. (2), for a variety of manifold geometries. Finding the maximum margin solution is challenging, as standard methods to solve SVM problems are limited to a finite number of input points. We have recently developed an efficient algorithm for finding the maximum margin solution in manifold classification and have used this method in the present work [see Ref. [22] and SM [16] (Sec. S5)].

IV. MANIFOLD GEOMETRY

A. Longitudinal and intrinsic coordinates

In this section, we address how the capacity to separate a set of manifolds can be related to their geometry, in particular, to their shape within the D-dimensional affine subspace. Since the projections of all points in a manifold onto the translation vector $\mathbf{c}^\mu$ are the same, $\vec{\mathbf{x}}^\mu \cdot \mathbf{c}^\mu = \|\mathbf{c}^\mu\|^2 = 1$, it is convenient to parametrize the $D+1$ affine

basis vectors such that $\mathbf{u}^{\mu}_{D+1} = \mathbf{c}^{\mu}$. In these coordinates, the $(D+1)$-dimensional vector representation of $\mathbf{c}^{\mu}$ is $\vec{C} = (0, 0, \ldots, 0, 1)$. This parametrization is convenient since it constrains the manifold variability to be in the first D components of $\vec{S}$, while the $D+1$ coordinate is a longitudinal variable measuring the distance of the manifold affine subspace from the origin. We can then write the $(D+1)$-dimensional vectors $\vec{T} = (\vec{t}, t_0)$, where $t_0 = \vec{T} \cdot \vec{C}$, $\vec{V} = (\vec{v}, v_0)$, $\vec{S} = (\vec{s}, 1)$, and $\tilde{S} = (\tilde{s}, 1)$, where lowercase vectors denote vectors in $\mathbb{R}^D$. We will also refer to the D-dimensional intrinsic vector $\tilde{s}$ as the anchor point. In this notation, the capacity, Eqs. (16) and (17), can be written as

$$\alpha_M^{-1}(\kappa) = \int D\vec{t} \int Dt_0 \frac{[-t_0 - \vec{t} \cdot \tilde{s}(\vec{t}, t_0) + \kappa]_+^2}{1 + \|\tilde{s}(\vec{t}, t_0)\|^2}. \quad (20)$$

It is clear from this form that when $D = 0$, or when all the vectors in the affine subspace obey $\tilde{s}(\vec{t}, t_0) = 0$, the capacity reduces to the Gardner result, Eq. (9). Since $\vec{V} \cdot \tilde{S} = \vec{v} \cdot \tilde{s} + v_0$, for all $\tilde{S}$, $g(\vec{V}) = g(\vec{v}) + v_0$, and the support function can be expressed as

$$g(\vec{v}) = \min_{\tilde{s}} \{\vec{v} \cdot \tilde{s} | \tilde{s} \in \mathcal{S}\}. \quad (21)$$

The KKT relations can be written as $\vec{v} = \vec{t} + \lambda\tilde{s}$, where $\lambda = v_0 - t_0 \geq 0$, $g(\vec{v}) \geq \kappa - v_0$, $\lambda[g(\vec{v}) + v_0 - \kappa] = 0$, and $\tilde{s}$ minimizes the overlap with $\vec{v}$. The resultant equation for λ (or v_0) is $\lambda = [-t_0 - \vec{t} \cdot \tilde{s}(\vec{t}, t_0) + \kappa]_+/(1 + \|\tilde{s}(\vec{t}, t_0)\|^2)$, which agrees with Eq. (17).

B. Types of supports

Using the above coordinates, we can elucidate the conditions for the different types of support manifolds defined in the previous section. To do this, we fix the random vector $\vec{t}$ and consider the qualitative change in the anchor point $\tilde{s}(\vec{t}, t_0)$ as t_0 decreases from $+\infty$ to $-\infty$.

Interior manifolds $(k = 0)$:—For sufficiently positive t_0, the manifold is interior to the margin plane, i.e., $\lambda = 0$ with corresponding support dimension $k = 0$. Although not contributing to the inverse capacity and solution vector, it is useful to associate anchor points to these manifolds defined as the closest point on the manifold to the margin plane: $\tilde{s}(\vec{t}) = \arg\min_{\tilde{s} \in \mathcal{S}} \vec{t} \cdot \tilde{s} = \nabla g(\vec{t})$, since $\vec{v} = \vec{t}$. This definition ensures continuity of the anchor point for all $\lambda \geq 0$.

This interior regime holds when $g(\vec{t}) > \kappa - t_0$ or, equivalently, for $t_0 - \kappa > t_{\text{touch}}(\vec{t})$, where

$$t_{\text{touch}}(\vec{t}) = -g(\vec{t}). \quad (22)$$

Nonzero contributions to the capacity only occur outside this interior regime when $g(\vec{t}) + t_0 - \kappa < 0$, in which case $\lambda > 0$. Thus, for all support dimensions $k > 0$, the solution for $\vec{v}$ is active, satisfying the equality condition, $g(\vec{v}) + v_0 - \kappa = 0$, so that, from Eq. (13),

$$g(\vec{v}) + t_0 - \kappa + \lambda = 0, \quad (23)$$

outside the interior regime.

Fully supporting manifolds $(k = D + 1)$:—When t_0 is sufficiently negative, $v_0 = \kappa$, $\vec{v} = 0$, and $\lambda = -t_0 + \kappa$. The anchor point $\tilde{s}(\vec{T})$, which obeys the KKT equations, resides in the interior of conv$(\mathcal{S})$,

$$\tilde{s}(\vec{t}, t_0) = \frac{-\vec{t}}{\kappa - t_0}. \quad (24)$$

For a fixed $\vec{t}$, t_0 must be negative enough, $t_0 - \kappa < t_{\text{fs}}$, where

$$t_{\text{fs}}(\vec{t}) = \arg\max\left\{t_0 | \frac{-\vec{t}}{\kappa - t_0} \in \text{conv}(\mathcal{S})\right\}, \quad (25)$$

guaranteeing that $\tilde{s}(\vec{t}, t_0) \in \text{conv}(\mathcal{S})$. The contribution of this regime to the capacity is

$$F(\vec{T}) = \|\vec{v} - \vec{t}\|^2 + (v_0 - t_0)^2 = \|\vec{t}\|^2 + (\kappa - t_0)^2; \quad (26)$$

see Eq. (12). Finally, for values of t_0 such that $t_{\text{fs}}(\vec{t}) \leq t_0 - \kappa \leq t_{\text{touch}}(\vec{t})$, the manifolds are partially supporting with support dimension $1 \leq k \leq D$. Examples of different supporting regimes are illustrated in Fig. 4.
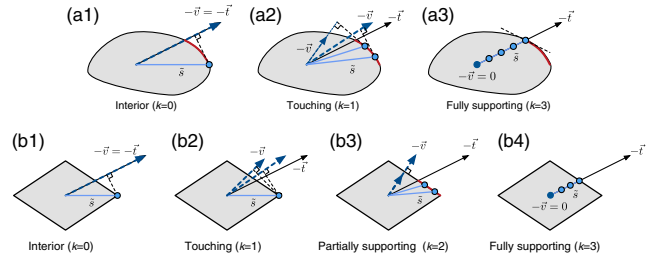

FIG. 4. Determining the anchor points of the Gaussian distributed vector $(\vec{t}, t_0)$ onto the convex hull of the manifold, denoted as $\tilde{s}(\vec{t}, t_0)$. Here, we show, for the same vector $\vec{t}$, the change in $\tilde{s}(\vec{t}, t_0)$ as t_0 decreases from $+\infty$ to $-\infty$. (a) $D = 2$ strictly convex manifold. For sufficiently positive t_0, the vector $-\vec{t}$ obeys the constraint $g(\vec{t}) + t_0 > \kappa$; hence, $-\vec{t} = -\vec{v}$, and the configuration corresponds to an interior manifold (support dimension $k = 0$). For intermediate values where $(t_{\text{touch}} > t_0 - \kappa > t_{\text{fs}})$, $(-\vec{t}, -t_0)$ violates the above constraints, and $\tilde{s}$ is a point on the boundary of the manifold that maximizes the projection on the manifold of a vector $(-\vec{v}, -v_0)$ that is the closest to $(-\vec{t}, -t_0)$ and obeys $g(\vec{v}) + t_0 + \lambda = \kappa$. Finally, for larger values of $-t_0$, $\vec{v} = 0$, and $\tilde{s}$ is a point in the interior of the manifold in the direction of $-\vec{t}$ (fully supporting with $k = 3$). (b) $D = 2$ square manifold. Here, in both the interior and touching regimes, $\tilde{s}$ is a vertex of the square. In the fully supporting regime, the anchor point is in the interior and collinear to $-\vec{t}$. There is also a partially supporting regime when $-t_0$ is slightly below t_{fs}. In this regime, $-\vec{v}$ is perpendicular to one of the edges and $\tilde{s}$ resides on an edge, corresponding to manifolds whose intersections with the margin planes are edges $(k = 2)$.

C. Effects of size and margin

We now discuss the effect of changing manifold size and the imposed margin on both capacity and geometry. As described by Eq. (4), a change in the manifold size corresponds to scaling every $\tilde{s}$ by a scalar r.

Small size:—If $r \to 0$, the manifold shrinks to a point (the center), whose capacity reduces to that of isolated points. However, in the case where $D \gg 1$, the capacity may be affected by the manifold structure even if $r \ll 1$; see Sec. IV F. Nevertheless, the underlying support structure is simple. When r is small, the manifolds have only two support configurations. For $t_0 - \kappa > 0$, the manifold is interior with $\lambda = 0$, $v_0 = t_0$, and $\vec{v} = \vec{t}$. When $t_0 - \kappa < 0$, the manifold becomes touching with support dimension $k = 1$. In that case, because $\tilde{s}$ has a small magnitude, $\vec{v} \approx \vec{t}$, $\lambda \approx \kappa - t_0$, and $v_0 \approx \kappa$. Thus, in both cases, $\vec{v}$ is close to the Gaussian vector $\vec{t}$. The probability of other configurations vanishes.

Large size:—In the large size limit $r \to \infty$, separating the manifolds is equivalent to separating the affine subspaces. As we show in Appendix C, when $r \gg 1$, there are two main support structures. With probability $H(-\kappa) = \int_{-\kappa}^{\infty} Dt_0$, the manifolds are fully supporting; namely, the underlying affine subspaces are parallel to the margin plane. This is the first regime corresponding to fully supporting manifolds, which contributes to the inverse capacity an amount $H(-\kappa)D + \alpha_0^{-1}(\kappa)$. The second regime corresponds to touching or partially supporting manifolds, such that the angle between the affine subspace and the margin plane is almost zero, and contributes an amount $H(\kappa)D$ to the inverse capacity. Combining the two contributions, we obtain, for large sizes, $\alpha_M^{-1} = D + \alpha_0^{-1}(\kappa)$, consistent with Eq. (8).

Large margin:—For a fixed $\vec{T}$, Eq. (25) implies that larger κ increases the probability of being in the supporting regimes. Increasing κ also shrinks the magnitude of $\tilde{s}$ according to Eq. (24). Hence, when $\kappa \gg 1$, the capacity becomes similar to that of P random points and the corresponding capacity is given by $\alpha_M \approx \kappa^{-2}$, independent of manifold geometry.

Manifold centers:—The theory of manifold classification described in Sec. III does not require the notion of a manifold center. However, once we understand how scaling the manifold sizes by a parameter r in Eq. (4) affects their capacity, the center points about which the manifolds are scaled need to be defined. For many geometries, the center is a point of symmetry such as for an ellipsoid. For general manifolds, a natural definition would be the center of mass of the anchor points $\tilde{S}(\vec{T})$ averaging over the Gaussian measure of $\vec{T}$. We will adopt here a simpler definition for the center provided by the Steiner point for convex bodies [23], $\vec{S}^0 = (\vec{s}^0, 1)$, with

$$\vec{s}^0 = \langle \nabla g_{\mathcal{S}}(\vec{t}) \rangle_{\vec{t}}, \qquad (27)$$

and the expectation is over the Gaussian measure of $\vec{t} \in \mathbb{R}^D$. This definition coincides with the center of mass

of the anchor points when the manifold size is small. Furthermore, it is natural to define the geometric properties of manifolds in terms of *centered manifolds* where the manifolds are shifted within their affine subspace so that the center and orthogonal translation vector coincide, i.e., $\vec{s}^0 = \vec{c}$ with $\|\vec{s}^0\| = \|\vec{c}\| = 1$. This means that all lengths are defined relative to the distance of the centers from the origin and the D-dimensional intrinsic vectors $\vec{s}$ give the offset relative to the manifold center.

D. Manifold anchor geometry

The capacity equation, Eq. (20), motivates defining geometrical measures of the manifolds, which we call "manifold anchor geometry." Manifold anchor geometry is based on the statistics of the anchor points $\tilde{s}(\vec{t}, t_0)$ induced by the Gaussian random vector $(\vec{t}, t_0)$, which are relevant to the capacity in Eq. (20). These statistics are sufficient for determining the classification properties and supporting structures associated with the maximum margin solution. We accordingly define the manifold anchor radius and dimension as follows.

Manifold anchor radius:—Denoted R_M, it is defined by the mean squared length of $\tilde{s}(\vec{t}, t_0)$,

$$R_M^2 = \langle \|\tilde{s}(\vec{t}, t_0)\|^2 \rangle_{\vec{t}, t_0}. \qquad (28)$$

Manifold anchor dimension: D_M is given by

$$D_M = \langle (\hat{t} \cdot \hat{s}(\vec{t}, t_0))^2 \rangle_{\vec{t}, t_0}, \qquad (29)$$

where $\hat{s}$ is a unit vector in the direction of $\tilde{s}$. The anchor dimension measures the angular spread between $\vec{t}$ and the corresponding anchor point $\tilde{s}$ in D dimensions. Note that the manifold dimension obeys $D_M \leq \langle \|\vec{t}\|^2 \rangle = D$. Whenever there is no ambiguity, we will call R_M and D_M the manifold radius and dimension, respectively.

These geometric descriptors offer a rich description of the manifold properties relevant for classification. Since $\vec{v}$ and $\tilde{s}$ depend in general on $t_0 - \kappa$, the above quantities are averaged not only over $\vec{t}$ but also over t_0. For the same reason, the manifold anchor geometry also depends upon the imposed margin.

E. Gaussian geometry

We have seen that, for small manifold sizes, $\vec{v} \approx \vec{t}$, and the anchor points $\tilde{s}$ can be approximated by $\tilde{s}(\vec{t}) = \nabla g_{\mathcal{S}}(\hat{t})$. Under these conditions, the geometry simplifies as shown in Fig. 5. For each Gaussian vector $\vec{t} \in \mathbb{R}^D$, $\tilde{s}(\vec{t})$ is the point on the manifold that first touches a hyperplane normal to $-\vec{t}$ as it is translated from infinity towards the manifold. Aside from a set of measure zero, this touching point is a unique point on the boundary of $\mathrm{conv}(\mathcal{S})$. This procedure is similar to that used to define the well-known Gaussian mean width, which in our notation equals $w(\mathcal{S}) = -2\langle g_{\mathcal{S}}(\vec{t}) \rangle_{\vec{t}}$ [24].
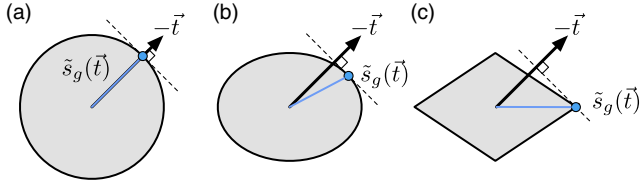
FIG. 5. Gaussian anchor points, with mapping from $\vec{t}$ to points $\tilde{s}_g(\vec{t})$, showing the relation between $-\vec{t}$ and the point on the manifold that touches a hyperplane orthogonal to $\vec{t}$. $D = 2$ manifolds shown are (a) circle, (b) ℓ_2 ellipsoid, (c) polytope manifold. In (c), only for values of $\vec{t}$ of measure zero (when $\vec{t}$ is exactly perpendicular to an edge) will $\tilde{s}_g(\vec{t})$ lie along the edge; otherwise it coincides with a vertex of the polytope. In all cases, $\tilde{s}_g(\vec{t})$ will be in the interior of the convex hulls only for $\vec{t} = 0$. Otherwise, it is restricted to their boundary.

Note that, for small sizes, the touching point does not depend on t_0 or κ, so its statistics are determined only by the shape of conv($\mathcal{S}$) relative to its center. This motivates defining a simpler manifold Gaussian geometry denoted by the subscript g, which highlights its dependence over a D-dimensional Gaussian measure for $\vec{t}$.

Gaussian radius:—Denoted by R_g, this measures the mean square amplitude of the Gaussian anchor point, $\tilde{s}_g(\vec{t}) = \nabla g_S(\hat{t})$:

$$R_g^2 = \langle \|\tilde{s}_g(\vec{t})\|^2 \rangle_{\vec{t}}, \tag{30}$$

where the expectation is over the D-dimensional Gaussian $\vec{t}$.

Gaussian dimension:—D_g is defined by

$$D_g = \langle (\vec{t} \cdot \hat{s}_g(\vec{t}))^2 \rangle_{\vec{t}}, \tag{31}$$

where $\hat{s}_g$ is the unit vector in the direction of $\tilde{s}_g$. While R_g^2 measures the total variance of the manifold, D_g measures the angular spread of the manifold, through the statistics of the angle between $\vec{t}$ and its Gaussian anchor point. Note that $D_g \leq \langle \|\vec{t}\|^2 \rangle = D$.

It is important to note that, even in this limit, our geometrical definitions are not equivalent to conventional geometrical measures such as the longest chord or second-order statistics induced from a uniform measure over the boundary of conv($\mathcal{S}$). For the special case of D-dimensional ℓ_2 balls with radius R, $\tilde{s}_g(\vec{t})$ is the point on the boundary of the ball in the direction of $\vec{t}$ so that $R_g = R$ and $D_g = D$. However, for general manifolds, D_g can be much smaller than the manifold affine dimension D, as illustrated in some examples later.

We recapitulate the essential difference between the Gaussian geometry and the full manifold anchor geometry. In the Gaussian case, the radius is an intrinsic property of the shape of the manifold in its affine subspace and is invariant to changing its distance from the origin. Thus,

scaling the manifold by a global scale factor r as defined in Eq. (4) results in scaling R_g by the factor r. Likewise, the dimensionality D_g is invariant to a global scaling of the manifold size. In contrast, the anchor geometry does not obey such invariance for larger manifolds. The reason is that the anchor point depends on the longitudinal degrees of freedom, namely, on the size of the manifold relative to the distance from the center. Hence, R_M need not scale linearly with r, and D_M will also depend on r. Thus, the anchor geometry can be viewed as describing the general relationship between the signal (center distance) and noise (manifold variability) in classification capacity. We also note that the manifold anchor geometry automatically accounts for the rich support structure described in Sec. IV D. In particular, as t_0 decreases, the statistics of the anchor points change from being concentrated on the boundary of conv($\mathcal{S}$) to its interior. Additionally, for manifolds that are not strictly convex and intermediate values of t_0, the anchor statistics become concentrated on k-dimensional facets of the convex hull, corresponding to partially supported manifolds.

We illustrate the difference between the two geometries in Fig. 6 with two simple examples: a $D = 2$ ℓ_2 ball and a $D = 2$ ℓ_2 ellipse. In both cases, we consider the distribution of $\|\tilde{s}\|$ and $\theta = \cos^{-1}(-\hat{t} \cdot \hat{s})$, the angle between $-\vec{t}$ and $\tilde{s}$. For the ball with radius r, the vectors $-\vec{t}$ and $\tilde{s}$ are parallel, so the angle is always zero. For the manifold anchor geometry, $\tilde{s}$ may lie inside the ball in the fully supporting region. Thus, the distribution of $\|\tilde{s}\|$ consists of a mixture of a delta function at r, corresponding to the interior and touching regions, and a smoothly varying distribution corresponding to the fully supporting region.

Figure 6 also shows the corresponding distributions for a two-dimensional ellipsoid with major and minor radius, $R_1 = r$ and $R_2 = \frac{1}{2}r$. For the Gaussian geometry, the distribution of $\|\tilde{s}\|$ has finite support between R_1 and R_2, whereas the manifold anchor geometry has support also below R_2. Since $\vec{t}$ and $\tilde{s}$ need not be parallel, the distribution
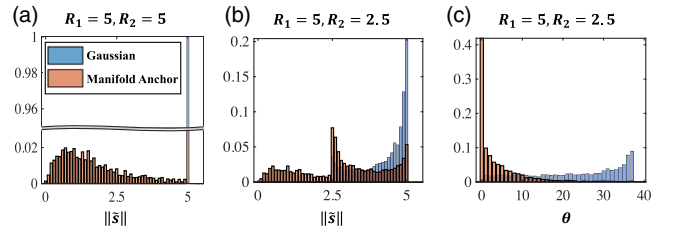

FIG. 6. Distribution of $\|\tilde{s}\|$ (norm of manifold anchor vectors) and $\theta = \angle(-\vec{t}, \tilde{s})$ for $D = 2$ ellipsoids. (a) Distribution of $\tilde{s}$ for an ℓ_2 ball with $R = 5$. The Gaussian geometry is peaked at $\|\tilde{s}\| = r$ (blue), while $\|\tilde{s}\| < r$ has nonzero probability for the manifold anchor geometry (red). (b),(c) $D = 2$ ellipsoids, with radii $R_1 = 5$ and $R_2 = \frac{1}{2}R_1$. (b) Distribution of $\|\tilde{s}\|$ for Gaussian geometry (blue) and anchor geometry (red). (c) Corresponding distribution of $\theta = \angle(-\vec{t}, \tilde{s})$.

of the angle varies between zero and $(\pi/4)$, with the manifold anchor geometry being more concentrated near zero due to contributions from the fully supporting regime. In Sec. VI, we will show how the Gaussian geometry becomes relevant even for larger manifolds when the labels are highly imbalanced, i.e., sparse classification.

F. Geometry and classification of high-dimensional manifolds

In general, linear classification as expressed in Eq. (20) depends on high-order statistics of the anchor vectors $\tilde{s}(\vec{t}, t_0)$. Surprisingly, our analysis shows that, for high-dimensional manifolds, the classification capacity can be described in terms of the statistics of R_M and D_M alone. This is particularly relevant as we expect that, in many applications, the affine dimension of the manifolds is large. More specifically, we define high-dimensional manifolds as manifolds where the manifold dimension is large, i.e., $D_M \gg 1$ (but still finite in the thermodynamic limit). In practice, we find that $D_M \gtrsim 5$ is sufficient. Our analysis below elucidates the interplay between size and dimension, namely, how small R_M needs to be for high-dimensional manifolds to have a substantial classification capacity.

In the high-dimensional regime, the mean-field equations simplify due to self-averaging of terms involving sums of components t_i and $\tilde{s}_i$. The quantity, $\vec{t} \cdot \tilde{s}$, that appears in the capacity, Eq. (20), can be approximated as $-\vec{t} \cdot \tilde{s} \approx \kappa_M$, where we introduce the effective manifold margin $\kappa_M = R_M\sqrt{D_M}$. Combined with $\|\tilde{s}\|^2 \approx R_M^2$, we obtain

$$\alpha_M(\kappa) \approx \alpha_0 \left(\frac{\kappa + \kappa_M}{\sqrt{1 + R_M^2}} \right), \tag{32}$$

where α_0 is the capacity for P random points, Eq. (9). To gain insight into this result, we note that the effective margin on the center is its mean distance from the point closest to the margin plane $\tilde{s}$, which is roughly the mean of $-\vec{t} \cdot \tilde{s} \approx R_M\sqrt{D_M}$. The denominator in Eq. (32) indicates that the margin needs to be scaled by the input norm, $\sqrt{1 + R_M^2}$. In Appendix B, we show that Eq. (32) can also be written as

$$\alpha_M(\kappa) \approx \alpha_{\text{Ball}}(\kappa, R_M, D_M), \tag{33}$$

namely, the classification capacity of a general high-dimensional manifold is well approximated by that of ℓ_2 balls with dimension D_M and radius R_M.

Scaling regime:—Equation (32) implies that, to obtain a finite capacity in the high-dimensional regime, the effective margin $\kappa_M = R_M\sqrt{D_M}$ needs to be of order unity, which requires the radius to be small, scaling for large D_M as $R_M = O(D_M^{-\frac{1}{2}})$. In this scaling regime, the calculation of the capacity and geometric properties are particularly simple. As argued above, when the radius is small, the components

of $\tilde{s}$ are small; hence $\vec{v} \approx \vec{t}$, and the Gaussian statistics for the geometry suffice. Thus, we can replace R_M and D_M in Eqs. (32) and (33) by R_g and D_g, respectively. Note that, in the scaling regime, the factor proportional to $1 + R_g^2$ in Eq. (32) is the next-order correction to the overall capacity, since R_g is small. Notably, the margin in this regime, $\kappa_g = R_g\sqrt{D_g}$, is equal to half the Gaussian mean width of convex bodies, $\kappa_g \approx \frac{1}{2}w(\mathcal{S})$ [25]. As for the support structure, since the manifold size is small, the only significant contributions arise from the interior ($k = 0$) and touching ($k = 1$) supports.

Beyond the scaling regime:—When R_g is not small, the anchor geometry, R_M and D_M, cannot be adequately described by the Gaussian statistics, R_g and D_g. In this case, the manifold margin $\kappa_M = R_M\sqrt{D_M}$ is large and Eq. (32) reduces to

$$\alpha_M \approx \frac{1 + R_M^{-2}}{D_M}, \tag{34}$$

where we have used $\alpha_0(\kappa) \approx \kappa^{-2}$ for large margins and assumed that $\kappa \ll \kappa_M$. For strictly convex high-dimensional manifolds with $R_M = O(1)$, only the touching regime ($k = 1$) contributes significantly to the geometry and, hence, to the capacity. For manifolds that are not strictly convex, partially supporting solutions with $k \ll D$ can also contribute to the capacity.

Finally, when R_M is large, fully supporting regimes with $k \approx D$ contribute to the geometry, in which case, the manifold anchor dimension approaches the affine dimension $D_M \rightarrow D$ and Eqs. (32) and (34) reduce to $\alpha_M \approx (1/D)$, as expected.

V. EXAMPLES

A. Strictly convex manifolds: ℓ_2 ellipsoids

The family of ℓ_2 ellipsoids are examples of manifolds that are strictly convex. Strictly convex manifolds have smooth boundaries that do not contain corners, edges, or flats (see Appendix B). Thus, the description of the anchor geometry and support structures is relatively simple. The reason is that the anchor vectors $\tilde{s}(\vec{t}, t_0)$ correspond to interior ($k = 0$), touching ($k = 1$), or fully supporting ($k = D + 1$), while partial support ($1 < k < D + 1$) is not possible. The ℓ_2 ellipsoid geometry can be solved analytically; nevertheless, because it has less symmetry than the sphere, it exhibits some salient properties of manifold geometry, such as nontrivial dimensionality and nonuniform measure of the anchor points.

We assume the ellipsoids are centered relative to their symmetry centers as described in Sec. IV, and they can be parametrized by the set of points $M^\mu = \mathbf{x}_0^\mu + \sum_{i=1}^D s_i \mathbf{u}_i^\mu$, where

$$S = \left\{ \vec{s} \Big| \sum_{i=1}^{D} \left(\frac{s_i}{R_i} \right)^2 \leq 1 \right\}. \tag{35}$$

The components of $\mathbf{u}_i^\mu$ and of the ellipsoid centers $\mathbf{x}_0^\mu$ are i.i.d. Gaussian distributed with zero mean and variance $(1/\sqrt{N})$ so that they are orthonormal in the large N limit. The radii R_i represent the principal radii of the ellipsoids relative to the center. The anchor points $\tilde{s}(\vec{t}, t_0)$ can be computed explicitly (details in the Appendix B), corresponding to three regimes.

The interior regime ($k = 0$) occurs when $t_0 - \kappa \geq t_{\text{touch}}(\vec{t})$, where

$$t_{\text{touch}}(\vec{t}) = \sqrt{\sum_{i=1}^{D} R_i^2 t_i^2}. \tag{36}$$

Here, $\lambda = 0$, resulting in zero contribution to the inverse capacity.

The touching regime ($k = 1$) holds when $t_{\text{touch}}(\vec{t}) > t_0 - \kappa > t_{\text{fs}}(\vec{t})$ and

$$t_{\text{fs}}(\vec{t}) = -\sqrt{\sum_{i=1}^{D} \left(\frac{t_i}{R_i} \right)^2}. \tag{37}$$

Finally, the fully supporting regime ($k = D + 1$) occurs when $t_{\text{fs}}(\vec{t}) > t_0 - \kappa$. The full expression for the capacity for ℓ_2 ellipsoids is given in Appendix B. Below, we focus on the interesting cases of ellipsoids with $D \gg 1$.

High-dimensional ellipsoids:—It is instructive to apply the general analysis of high-dimensional manifolds to ellipsoids with $D \gg 1$. We will distinguish among different-size regimes by assuming that all the radii of the ellipsoid are scaled by a global factor r. In the high-dimensional regime, due to self-averaging, the boundaries of the touching and fully supporting transitions can be approximated by

$$t_{\text{touch}} = \sqrt{\sum_{i=1}^{D} R_i^2} \tag{38}$$

$$t_{\text{fs}} = -\sqrt{\sum_{i=1}^{D} R_i^{-2}}, \tag{39}$$

both independent of $\vec{t}$. Then, as long as $R_i \sqrt{D}$ are not large (see below), $t_{\text{fs}} \to -\infty$, and the probability of fully supporting vanishes. Discounting the fully supporting regime, the anchor vectors $\tilde{s}(\vec{t}, t_0)$ are given by

$$\tilde{s}_i = -\lambda_i t_i \tag{40}$$

$$\lambda_i = \frac{R_i^2}{Z(1 + R_i^2)}, \tag{41}$$

where the normalization factor $Z^2 = \sum_{i=1}^{D} [R_i^2/(1 + R_i^2)^2]$ (see Appendix B). The capacity for high-dimensional ellipsoids is determined via Eq. (32) with manifold anchor radius

$$R_M^2 = \sum_{i=1}^{D} \lambda_i^2 \tag{42}$$

and anchor dimension

$$D_M = \frac{(\sum_{i=1}^{D} \lambda_i)^2}{\sum_{i=1}^{D} \lambda_i^2}. \tag{43}$$

Note that R_M/r and D_M are not invariant to scaling the ellipsoid by a global factor r, reflecting the role of the fixed centers.

The anchor covariance matrix:—We can also compute the covariance matrix of the anchor points. This matrix is diagonal in the principal directions of the ellipsoid, and its eigenvalues are λ_i^2. It is interesting to compare D_M with a well-known measure of an effective dimension of a covariance matrix, the participation ratio [26,27]. Given a spectrum of eigenvalues of the covariance matrix, in our notation λ_i^2, we can define a generalized participation ratio as $PR_q = [(\sum_{i=1}^{D} \lambda_i^q)^2 / \sum_{i=1}^{D} \lambda_i^{2q}]$, $q > 0$. The conventional participation ratio uses $q = 2$, whereas D_M uses $q = 1$.

Scaling regime:—In the scaling regime where the radii are small, the radius and dimension are equivalent to the Gaussian geometry:

$$R_g^2 = \frac{\sum_{i=1}^{D} R_i^4}{\sum_{i=1}^{D} R_i^2} \tag{44}$$

$$D_g = \frac{(\sum_{i=1}^{D} R_i^2)^2}{\sum_{i=1}^{D} R_i^4}, \tag{45}$$

and the effective margin is given as

$$\kappa_g = \left(\sum_{i=1}^{D} R_i^2 \right)^{1/2}. \tag{46}$$

As can be seen, D_g and (R_g/r) are invariant to scaling all the radii by r, as expected. It is interesting to compare the above Gaussian geometric parameters with the statistics induced by a uniform measure on the surface of the ellipsoid. In this case, the covariance matrix has eigenvalues $\lambda_i^2 = R_i^2$, and the total variance is equal to $\sum_i R_i^2$. In contrast, in the Gaussian geometry, the eigenvalues of the covariance matrix λ_i^2 are proportional to R_i^4. This and the

corresponding expression (44) for the squared radius is the result of a nonuniform induced measure on the surface of the ellipse in the anchor geometry, even in its Gaussian limit.

Beyond the scaling regime:—When $R_i = O(1)$, the high-dimensional ellipsoids all become touching manifolds since $t_{touch} \rightarrow -\infty$ and $t_{fs} \rightarrow \infty$. The capacity is small because $\kappa_M \gg 1$ and is given by Eq. (34) with Eqs. (42) and (43). Finally, when all $R_i \gg 1$, we have

$$R_M^2 = \frac{D}{\sum_{i=1}^{D} R_i^{-2}} = \frac{1}{\langle R_i^{-2} \rangle_i}. \qquad (47)$$

Here, R_M scales with the global scaling factor r, $D_M = D$, and $\alpha^{-1} = D$. Although the manifolds are touching, their angle relative to the margin plane is near zero. When $R_i \approx \sqrt{D}$ or larger, the fully supporting transition t_{fs} becomes order one, and the probability of fully supporting is significant.

In Fig. 7, we illustrate the behavior of high-D-dimensional ellipsoids, using ellipsoids with a bimodal distribution of their principal radii R_i: $R_i = r$, for $1 \leq i \leq 10$, and $R_i = 0.1r$, for $11 \leq i \leq 200$ [Fig. 7(a)]. Their properties are shown as a function of the overall scale r. Figure 7(b) shows numerical simulations of the capacity, the full mean-field solution, and the spherical high-dimensional approximation (with R_M and D_M). These calculations are all in good agreement, showing the accuracy of the mean-field theory and spherical approximation. As seen in Figs. 7(b)–7(d), the system is in the scaling regime for $r < 0.3$. In this regime, the manifold dimension is constant and equals $D_g \approx 14$, as predicted by the participation ratio, Eq. (45), and the manifold radius R_g is linear with r, as expected from Eq. (44). The ratio $(R_g/r) \approx 0.9$ is close to unity, indicating that, in the scaling regime, the system is dominated by the largest radii. For $r > 0.3$, the effective margin is larger than unity, and the system becomes increasingly affected by the full affine dimensionality of the ellipsoid, as seen by the marked increase in dimension, as well as a corresponding decrease in (R_M/r). For larger r, D_M approaches D and $\alpha_M^{-1} = D$. Figures 7(e1)–7(e3) show the distributions of the support dimension $0 \leq k \leq D+1$. In the scaling regime, the interior and touching regimes each have probability close to $\frac{1}{2}$, and the fully supporting regime is negligible. As r increases beyond the scaling regime, the interior probability decreases, and the solution is almost exclusively in the touching regime. For very high values of r, the fully supporting solution gains a substantial probability. Note that the capacity decreases to approximately $\frac{1}{D}$ for a value of r below which a substantial fraction of solutions are fully supporting. In this case, the touching ellipsoids all have a very small angle with the margin plane.

Until now, we have assumed that the manifold affine subspace dimension D is finite in the limit of large ambient
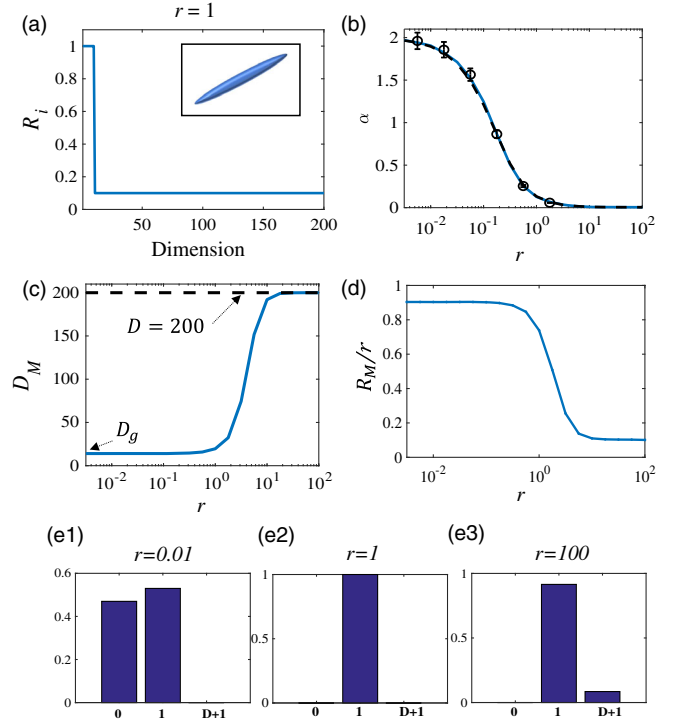

FIG. 7. Bimodal ℓ_2 ellipsoids. (a) Ellipsoidal radii R_i with $r = 1$. (b) Classification capacity as a function of scaling factor r: full mean-field theory capacity (blue lines); approximation of the capacity given by equivalent ball with R_M and D_M (black dashed lines); simulation capacity (circles), averaged over five repetitions, measured with 50 dichotomies per repetition. (c) Manifold dimension D_M as a function of r. (d) Manifold radius R_M relative to r. (e) Fraction of manifolds with support dimension k for different values of r: $k = 0$ (interior), $k = 1$ (touching), $k = D + 1$ (fully supporting). (e1) Small $r = 0.01$, where most manifolds are interior or touching. (e2) $r = 1$, where most manifolds are in the touching regime. (e3) $r = 100$, where the fraction of fully supporting manifolds is 0.085, predicted by $H(-t_{fs}) = H(\sqrt{\sum_i R_i^{-2}})$ [Eq. (39)].

dimension N. In realistic data, it is likely that the data manifolds are technically full rank, i.e., $D + 1 = N$, raising the question of whether our mean-field theory is still valid in such cases. We investigate this scenario by computing the capacity on ellipsoids containing a realistic distribution of radii. We have taken, as examples, a class of images from the ImageNet data set [28], and analyzed the SVD spectrum of the representations of these images in the last layer of a deep convolutional network, GoogLeNet [29]. The computed radii are shown in Fig. 8(a) and yield a value $D = N - 1 = 1023$. In order to explore the properties of such manifolds, we have scaled all the radii by an overall factor r in our analysis. Because of the decay in the distribution of radii, the Gaussian dimension for the ellipsoid is only about $D_g \approx 15$, much smaller than D or N, implying that, for small r, the manifolds are effectively low dimensional and the geometry is dominated by a small number of radii. As r increases above $r \gtrsim 0.03$,
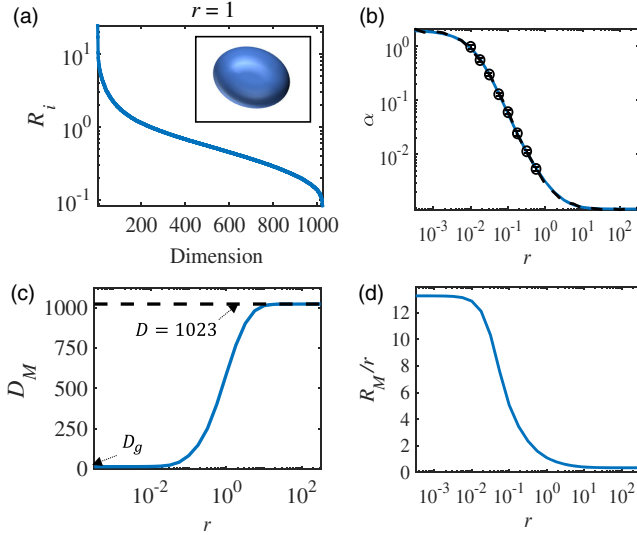
FIG. 8. Ellipsoids with radii computed from realistic image data. (a) SVD spectrum R_i, taken from the readout layer of GoogLeNet from a class of ImageNet images with $N = 1024$ and $D = 1023$. The radii are scaled by a factor r: $R'_i = rR_i$ for (b) Classification capacity as a function of r: full mean-field theory capacity (blue lines); approximation of the capacity as that of a ball with R_M and D_M from the theory for ellipsoids (black dashed lines); simulation capacity (circles), averaged over 5 repetitions, measured with 50 random dichotomies per each repetition. (c) Manifold dimension as a function of r. (d) Manifold radius relative to the scaling factor r, R_M/r as a function of r.

κ_M becomes larger than 1 and the solution leaves the scaling regime, resulting in a rapid increase in D_M and a rapid falloff in capacity, as shown in Figs. 8(b) and 8(c). Finally, for $r > 10$, we have $\alpha_M^{-1} \approx D_M \approx D$ approaching the lower bound for capacity as expected. The agreement between the numerical simulations and the mean-field estimates of the capacity illustrates the relevance of the theory for realistic data manifolds that can be full rank.

B. Convex polytopes: ℓ_1 ellipsoids

The family of ℓ_2 ellipsoids represents manifolds that are smooth and strictly convex. On the other hand, there are other types of manifolds whose convex hulls are not strictly convex. In this section, we consider D-dimensional ℓ_1 ellipsoids. They are prototypical of convex polytopes formed by the convex hulls of finite numbers of points in $\mathbb{R}^N$. The D-dimensional ℓ_1 ellipsoid is parametrized by radii $\{R_i\}$ and specified by the convex set $M^\mu = \mathbf{x}_0^\mu + \sum_{i=1}^D s_i \mathbf{u}_i^\mu$, with

$$\mathcal{S} = \left\{ \vec{s} \,\middle|\, \sum_{i=1}^D \frac{|s^i|}{R_i} \leq 1 \right\}. \tag{48}$$

Each manifold $M^\mu \in \mathbb{R}^N$ is centered at $\mathbf{x}_0^\mu$ and consists of a convex polytope with a finite number ($2D$) of vertices:

$\{\mathbf{x}_0^\mu \pm R_k \mathbf{u}_k^\mu, k = 1, ..., D\}$. The vectors $\mathbf{u}_i^\mu$ specify the principal axes of the ℓ_1 ellipsoids. For simplicity, we consider the case of ℓ_1 balls when all the radii are equal: $R_i = R$. We will concentrate on the cases when ℓ_1 balls are high dimensional; the case for ℓ_1 balls with $D = 2$ was briefly described in Ref. [15]. The analytical expression of the capacity is complex due to the presence of contributions from all types of supports, $1 \leq k \leq D + 1$. We address important aspects of the high-dimensional solution below.

High-dimensional ℓ_1 balls, scaling regime:—In the scaling regime, we have $\vec{v} \approx \vec{t}$. In this case, we can write the solution for the subgradient as

$$\tilde{s}_i(\vec{t}) = \begin{cases} -R\,\mathrm{sign}(t_i), & |t_i| > |t_j| \;\forall\; j \neq i \\ 0 & \text{otherwise} \end{cases}. \tag{49}$$

In other words, $\tilde{s}(\vec{t})$ is a vertex of the polytope corresponding to the component of $\vec{t}$ with the largest magnitude; see Fig. 5(c). The components of $\vec{t}$ are i.i.d. Gaussian random variables and, for large D, its maximum component $t_{\max}$ is concentrated around $\sqrt{2 \log D}$. Hence, $D_g = \langle (\hat{s} \cdot \vec{t})^2 \rangle = \langle t_{\max}^2 \rangle = 2 \log D$, which is much smaller than D. This result is consistent with the fact that the Gaussian mean width of a D-dimensional ℓ_1 ball scales with $\sqrt{\log D}$ and not with D [25]. Since all the points have norm R, we have $R_g = R$, and the effective margin is then given by $\kappa_g = R\sqrt{2 \log D}$, which is order unity in the scaling regime. In this regime, the capacity is given by simple relation $\alpha_M = \alpha_0([\kappa + R\sqrt{2 \log D}]/\sqrt{1 + R^2})$.

High-dimensional ℓ_1 balls, $R = O(1)$:—When the radius R is small as in the scaling regime, the only contributing solution is the touching solution ($k = 1$). When R increases, solutions with all values of k, $1 \leq k \leq D + 1$, occur, and the support can be any face of the convex polytope with dimension k. As R increases, the probability distribution $p(k)$ over k of the solution shifts to larger values. Finally, for large R, only two regimes dominate: fully supporting ($k = D + 1$) with probability $H(-\kappa)$ and partially supporting with $k = D$ with probability $H(\kappa)$.

We illustrate the behavior of ℓ_1 balls with radius r and affine dimension $D = 100$. In Fig. 9, (a) shows the linear classification capacity as a function of r. When $r \to 0$, the manifold approaches the capacity for isolated points, $\alpha_M \approx 2$, and when $r \to \infty$, $\alpha_M \approx (1/D) = 0.01$. The numerical simulations demonstrate that, despite the different geometry, the capacity of the ℓ_1 polytope is similar to that of a ℓ_2 ball with radius R_M and dimension D_M. In (b), for the scaling regime when $r < 0.1 = (1/\sqrt{D})$, we see $R_M \approx r$, but when $r \gg 1$, R_M is much smaller than r, despite the fact that all R_i of the polytope are equal. This is because when r is not small, the various faces and eventually interior of the polytope contribute to the anchor geometry. In (c), we see $D_M \approx 2 \log D$ in the scaling
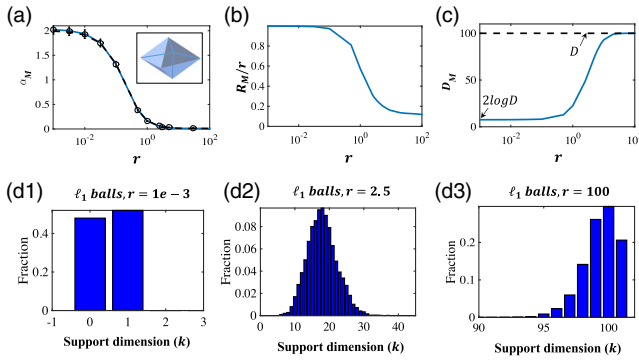
FIG. 9. Separability of ℓ_1 balls. (a) Linear classification capacity of ℓ_1 balls as a function of radius r with $D = 100$ and $N = 200$: MFT solution (blue lines), spherical approximation (black dashed lines), full numerical simulations (circles). Inset: Illustration of ℓ_1 ball. (b) Manifold radius R_M relative to the actual radius r. (c) Manifold dimension D_M as a function of r. In the small r limit, D_M is approximately $2 \log D$, while, in large r, D_M is close to D, showing how the solution is orthogonal to the manifolds when their sizes are large. (d1)–(d3) Distribution of support dimensions: (d1) $r = 0.001$, where most manifolds are either interior or touching; (d2) $r = 2.5$, where the support dimension has a peaked distribution; (d3) $r = 100$, where most manifolds are close to being fully supporting.

regime, while $D_M \to D$ as $r \to \infty$. In terms of the support structures, when $r = 0.001$ in the scaling regime (d1), most manifolds are either interior or touching. For intermediate sizes (d2), the support dimension is peaked at an intermediate value, and, finally, for very large manifolds (d3), most polytope manifolds are nearly fully supporting.

C. Smooth nonconvex manifolds: Ring manifolds

Many neuroscience experiments measure the responses of neuronal populations to a continuously varying stimulus, with one or a small number of degrees of freedom. A prototypical example is the response of neurons in visual cortical areas to the orientation (or direction of movement) of an object. When a population of neurons responds both to an object identity and to a continuous physical variation, the result is a set of smooth manifolds, each parametrized by a single variable, denoted θ, describing continuous curves in $\mathbb{R}^N$. Since, in general, the neural responses are not linear in θ, the curve spans more than one linear dimension. This smooth curve is not convex and is endowed with a complex nonsmooth convex hull. It is, thus, interesting to consider the consequences of our theory on the separability of smooth but nonconvex curves.

The simplest example considered here is the case where θ corresponds to a periodic angular variable such as the orientation of an image, and we call the resulting nonconvex curve a *ring* manifold. We model the neuronal responses as smooth periodic functions of θ, which can be parametrized by decomposing the neuronal responses in Fourier modes. Here, for each object,

$M^\mu = \mathbf{x}_0^\mu + \sum_{i=1}^D s_i \mathbf{u}_i^\mu$, where $\mathbf{x}_0^\mu$ represents the mean (over θ) of the population response to the object. The different D components correspond to the different Fourier components of the angular response, so that

$$s^{2n}(\theta) = \frac{R_n}{\sqrt{2}} \cos(n\theta)$$

$$s^{2n-1}(\theta) = \frac{R_n}{\sqrt{2}} \sin(n\theta), \qquad (50)$$

where R_n is the magnitude of the nth Fourier component for $1 \leq n \leq \frac{D}{2}$. The neural responses in Eq. (1) are determined by projecting onto the basis:

$$\mathbf{u}_{2n}^{\mu i} = \sqrt{2} \cos(n\theta^{\mu i})$$

$$\mathbf{u}_{2n-1}^{\mu i} = \sqrt{2} \sin(n\theta^{\mu i}). \qquad (51)$$

The parameters $\theta^{\mu i}$ are the preferred orientation angles for the corresponding neurons and are assumed to be evenly distributed between $-\pi \leq \theta_i^\mu \leq \pi$ (for simplicity, we assume that the orientation tunings of the neurons are all of the same shape and are symmetric around the preferred angle). The statistical assumptions of our analysis assume that the different manifolds are randomly positioned and oriented with respect to the others. For the ring manifold model, this implies that the mean responses $\mathbf{x}_0^\mu$ are independent random Gaussian vectors and also that the preferred orientation $\theta^{\mu i}$ angles are uncorrelated. With this definition, all the vectors $\vec{s} \in \mathcal{S}$ obey the normalization $\|\vec{s}\| = r$, where $r^2 = \sum_{n=1}^{\frac{D}{2}} R_n^2$. Thus, for each object, the ring manifold is a closed nonintersecting smooth curve residing on the surface of a D-dimensional sphere with radius r.

For the simplest case with $D = 2$, the ring manifold is equivalent to a circle in two dimensions. However, for larger D, the manifold is not convex and its convex hull is composed of faces with varying dimensions. In Fig. 10, we investigate the geometrical properties of these manifolds relevant for classification as a function of the overall scale factor r, where for simplicity we have chosen $R_n = r$ for all n. The most striking feature is the small dimension in the scaling regime, scaling roughly as $D_g \approx 2 \log D$. This logarithmic dependence is similar to that of the ℓ_1 ball polytopes. Then, as r increases, D_M increases dramatically from $2 \log D$ to D.

The similarity of the ring manifold to a convex polytope is also seen in the support dimension k of the manifolds. Support faces of dimension $k \neq 0$, 1, $D + 1$ are seen, implying the presence of partially supporting solutions. Interestingly, $(D/2) < k \leq D$ are excluded, indicating that the maximal face dimension of the convex hull is $(D/2)$. Each face is the convex hull of a set of $k \leq (D/2)$ points, where each point resides in the $2D$ subspace spanned by a
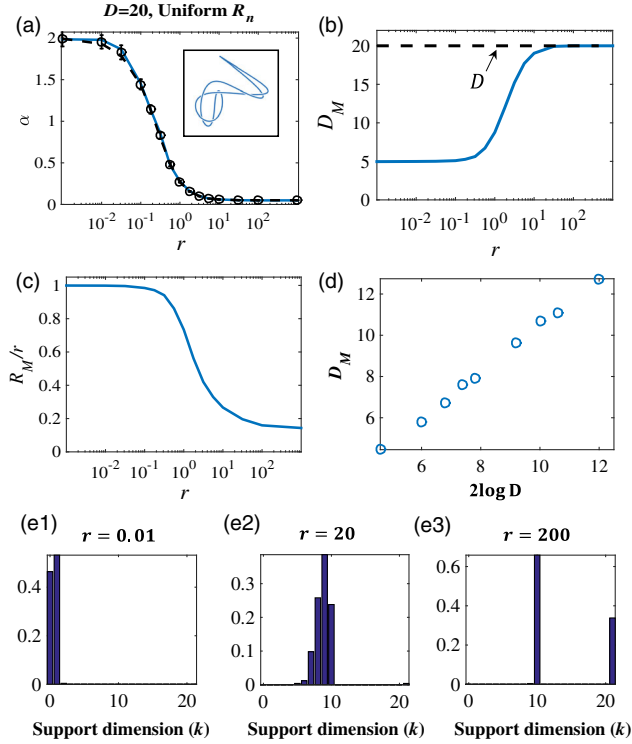
FIG. 10. Linear classification of D-dimensional ring manifolds, with uniform $R_n = \sqrt{(2/D)}\,r$. (a) Classification capacity for $D = 20$ as a function of r with $m = 200$ test samples (for details of numerical simulations, see SM [16], Sec. S9): mean-field theory (blue lines), spherical approximation (black dashed lines), numerical simulations (black circles). Inset: Illustration of a ring manifold. (b) Manifold dimension D_M, which shows $D_M \sim D$ in the large r limit, showing the orthogonality of the solution. (c) Manifold radius relative to the scaling factor r, (R_M/r), as a function of r. The fact that (R_M/r) becomes small implies that the manifolds are "fully supporting" to the hyperplane, showing the small radius structure. (d) Manifold dimension grows with affine dimension, D, as $2\log D$ for small $r = (1/2\sqrt{2\log D})$ in the scaling regime. (e1)–(e3) Distribution of support dimensions. (d1) $r = 0.01$, where most manifolds are either interior or touching; (d2) $r = 20$, where the support dimension has a peaked distribution, truncated at $(D/2)$; (d3) $r = 200$, where most support dimensions are $(D/2)$ or fully supporting at $D + 1$.

pair of (real and imaginary) Fourier harmonics. The ring manifolds are closely related to the trigonometric moment curve, whose convex hull geometrical properties have been extensively studied [30,31].

In conclusion, the smoothness of the convex hulls becomes apparent in the distinct patterns of the support dimensions [compare Figs. 8(d) and 10(e)]. However, we see that all manifolds with $D_g \sim 5$ or larger share common trends. As the size of the manifold increases, the capacity and geometry vary smoothly, exhibiting a smooth crossover from a high capacity with low radius and dimension to a low capacity with large radius and dimension. This crossover occurs as $R_g \propto 1/\sqrt{D_g}$. Also, our examples

demonstrate that, in many cases, when the size is smaller than the crossover value, the manifold dimensionality is substantially smaller than that expected from naive second-order statistics, highlighting the saliency and significance of our anchor geometry.

VI. MANIFOLDS WITH SPARSE LABELS

So far, we have assumed that the number of manifolds with positive labels is approximately equal to the number of manifolds with negative labels. In this section, we consider the case where the two classes are unbalanced such that the number of positively labeled manifolds is far less than the negatively labeled manifolds (the opposite scenario is equivalent). This is a special case of the problem of classification of manifolds with heterogeneous statistics, where manifolds have different geometries or label statistics. We begin by addressing the capacity of mixtures of manifolds and then focus on sparsely labeled manifolds.

A. Mixtures of manifold geometries

Our theory of manifold classification is readily extended to a heterogeneous ensemble of manifolds, consisting of L distinct classes. In the replica theory, the shape of the manifolds appear only in the free energy term G_1 [see Appendix, Eq. (A13)]. For a mixture statistics, the combined free energy is given by simply averaging the individual free energy terms over each class l. Recall that this free energy term determines the capacity for each shape, giving an individual inverse critical load α_l^{-1}. The inverse capacity of the heterogeneous mixture is then

$$\alpha^{-1} = \langle \alpha_l^{-1} \rangle_l, \tag{52}$$

where the average is over the fractional proportions of the different manifold classes. This remarkably simple but generic theoretical result enables analyzing diverse manifold classification problems, consisting of mixtures of manifolds with varying dimensions, shapes, and sizes.

Equation (52) is adequate for classes that differ in their geometry, independent of the assigned labels. In general, classes may differ in the label statistics (as in the sparse case studied below) or in geometry that is correlated with labels. For instance, the positively labeled manifolds may consist of one geometry, and the negatively labeled manifolds may have a different geometry. How do structural differences between the two classes affect the capacity of the linear classification? A linear classifier can take advantage of these correlations by adding a nonzero bias. Previously, it was assumed that the optimal separating hyperplane passes through the origin; this is reasonable when the two classes are statistically the same. However, when there are statistical differences between the label assignments of the two classes, Eq. (2) should be replaced by $y^\mu(\mathbf{w} \cdot \mathbf{x}^\mu - b) \geq \kappa$, where the bias b is chosen to maximize the mixture

capacity, Eq. (52). The effect of optimizing the bias is discussed in detail in the next section for the sparse labels and in other scenarios in the SM [16] (Sec. S3).

B. Manifolds with sparse labels

We define the sparsity parameter f as the fraction of positively labeled manifolds so that $f = 0.5$ corresponds to having balanced labels. From the theory of the classification of a finite set of random points, it is known that having sparse labels with $f \ll 0.5$ can drastically increase the capacity [12]. In this section, we investigate how the sparsity of manifold labels improves the manifold classification capacity.

If the separating hyperplane is constrained to go through the origin and the distribution of inputs is symmetric around the origin, the labeling y^μ is immaterial to the capacity. Thus, the effect of sparse labels is closely tied to having a nonzero bias. We thus consider inequality constraints of the form $y^\mu(\mathbf{w} \cdot \mathbf{x}^\mu - b) \geq \kappa$, and we define the bias-dependent capacity of general manifolds with label sparsity f, margin κ, and bias b, as $\alpha_M(\kappa, f, b)$. Next, we observe that the bias acts as a positive contribution to the margin for the positively labeled population and as a negative contribution to the negatively labeled population. Thus, the dependence of $\alpha_M(\kappa, f, b)$ on both f and b can be expressed as

$$\alpha_M^{-1}(\kappa, f, b) \equiv f\alpha_M^{-1}(\kappa + b) + (1 - f)\alpha_M^{-1}(\kappa - b), \quad (53)$$

where $\alpha_M(x)$ is the classification capacity with zero bias (and, hence, equivalent to the capacity with $f = 0.5$) for the same manifolds. Note that Eq. (53) is similar to Eq. (52) for mixtures of manifolds. The actual capacity with sparse labels is given by optimizing the above expression with respect to b, i.e.,

$$\alpha_M(\kappa, f) = \max_b \alpha_M(\kappa, f, b). \quad (54)$$

In the following, we consider for simplicity the effect of sparsity for zero margin, $\kappa = 0$.

Importantly, if D is not large, the effect of the manifold geometry in sparsely labeled manifolds can be much larger than that for nonsparse labels. For nonsparse labels, the capacity ranges between 2 and $(D + 0.5)^{-1}$, while for sparse manifolds, the upper bound can be much larger. Indeed, small-sized manifolds are expected to have capacity that increases upon decreasing f as $\alpha_M(0, f) \propto (1/f|\log f|)$, similar to P uncorrelated points [12]. This potential increase in the capacity of sparse labels is, however, strongly constrained by the manifold size, since, when the manifolds are large, the solution has to be orthogonal to the manifold directions so that $\alpha_M(0, f) \approx (1/D)$. Thus, the geometry of the manifolds plays an important role in controlling the effect of sparse labels on capacity. These aspects are already

seen in the case of sparsely labeled ℓ_2 balls (Appendix E). Here, we summarize the main results for general manifolds.

Sparsity and size:—There is a complex interplay between label sparsity and manifold size. Our analysis yields three qualitatively different regimes.

Low R_g:—When the Gaussian radius of the manifolds are small, i.e., $R_g < 1$, the extent of the manifolds is noticeable only when the dimension is high. Similar to our previous analysis of high-dimensional manifolds, we find here that the sparse capacity is equivalent to the capacity of sparsely labeled random points, with an effective margin given by $R_g\sqrt{D_g}$,

$$\alpha_M(f) \approx \alpha_0(f, \kappa = R_g\sqrt{D_g}), \quad (55)$$

where $\alpha_0(\kappa, f) = \max_b \alpha_0(\kappa, f, b)$ from the Gardner theory. It should be noted that κ_g in the above equation has a noticeable effect only for moderate sparsity. It has a negligible effect when $f \to 0$, since the bias is large and dominates over the margin.

Moderate sizes, $R_g > 2$:—In this case, the equivalence to the capacity of points breaks down. Remarkably, we find that the capacity of general manifolds with substantial size is well approximated by that of equivalent ℓ_2 balls with the same sparsity f and with dimension and radius equal to the Gaussian dimension and radius of the manifolds, namely,

$$\alpha_M(f) \approx \alpha_{\text{Ball}}(f, R_g, D_g). \quad (56)$$

Surprisingly, unlike the nonsparse approximation, where the equivalence of general manifold to balls, Eq. (33), is valid only for high-dimensional manifolds, in the sparse limit, the spherical approximation is not restricted to large D. Another interesting result is that the relevant statistics are given by the Gaussian geometry, R_g and D_g, even when R_g is not small. The reason is that, for small f, the bias is large. In that case, the positively labeled manifolds have large positive margin b and are fully supporting giving a contribution to the inverse capacity of b^2, regardless of their detailed geometry. On the other hand, the negatively labeled manifolds have large negative margin, implying that most of them are far from the separating plane (interior) and a small fraction have touching support. The fully supporting configurations have negligible probability; hence, the overall geometry is well approximated by the Gaussian quantities D_g and R_g.

Scaling relationship between sparsity and size:—A further analysis of the capacity of balls with sparse labels shows it retains a simple form $\alpha(\bar{f}, D)$ [see Appendix E, Eq. (E10)], which depends on R and f only through the scaled sparsity $\bar{f} = fR^2$. The reason for the scaling of f with R^2 is as follows. When the labels are sparse, the dominant contribution to the inverse capacity comes from the minority class, and so the capacity is $\alpha_{\text{Ball}}^{-1} \approx fb^2$ for large b. On the

other hand, the optimal value of b depends on the balance between the contributions from both classes and scales linearly with R, as it needs to overcome the local fields from the spheres. Thus, $\alpha^{-1} \propto fR^2$.

Combining Eq. (E10) with Eq. (56) yields, for general sparsely labeled manifolds,

$$\alpha_M^{-1} = \bar{f}\bar{b}^2 + \int_{\bar{b}}^{\infty} dt \chi_{D_g}(t)(t - \bar{b})^2, \qquad (57)$$

where scaled sparsity is $\bar{f} = f(1 + R_g^2) \lesssim 1$.

Note that we have defined the scaled sparsity $\bar{f} = f(1 + R_g^2)$ rather than fR^2 to yield a smoother crossover to the small R_g regime. Similarly, we define the optimal scaled bias as $\bar{b} = b\sqrt{1 + R_g^2} \gg 1$. Qualitatively, the function α is roughly proportional to $\bar{f}^{-1}$ with a proportionality constant that depends on D_g. In the extreme limit of $|\log \bar{f}| \gg D_g$, we obtain the sparse limit $\alpha \propto |\bar{f}\log \bar{f}|^{-1}$. Note that, if f is sufficiently small, the gain in capacity due to sparsity occurs even for large manifolds as long as $\bar{f} < 1$.

Large R_g regime, $\bar{f} > 1$:—Finally, when R_g is sufficiently large such that $\bar{f}$ increases above 1, $\bar{b}$ is of order 1 or smaller, and the capacity is small with a value that depends on the detailed geometry of the manifold. In particular, when $\bar{f} \gg 1$, the capacity of the manifold approaches $\alpha_M(f) \rightarrow (1/D)$, not $(1/D_g)$.

To demonstrate these remarkable predictions, Eq. (56), we present in Fig. 11 the capacity of three classes of sparsely labeled manifolds: ℓ_2 balls, ℓ_1 ellipsoids, and ring manifolds. In all cases, we show the results of numerical simulations of the capacity, the full mean-field solution, and the spherical approximation, Eq. (56), across several orders of magnitude of sparsity and size as a function of the
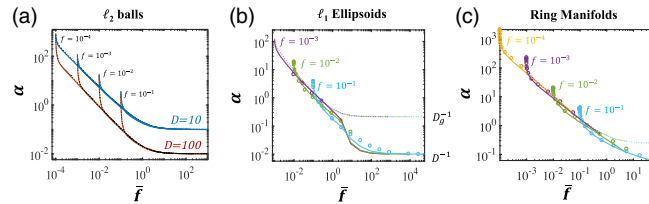
scaled sparsity $\bar{f} = f(1 + R^2)$. In each example, there is a good agreement between the three calculations for the range of $\bar{f} < 1$. Furthermore, the drop in α with increasing $\bar{f}$ is similar in all cases, except for an overall vertical shift, which is due to the different D_g, similar to the effect of dimension in ℓ_2 balls [Fig. 11(a)]. In the regime of moderate radii, results for different f and r all fall on a universal curve, which depends only on $\bar{f}$, as predicted by the theory. For small r, the capacity deviates from this scaling, as it is dominated by f alone, similar to sparsely labeled points. When $\bar{f} > 1$, the curves deviate from the spherical approximation. The true capacity (as revealed by simulations and full mean field) rapidly decreases with $\bar{f}$ and saturates at $(1/D)$, rather than to the $(1/D_g)$ limit of the spherical approximation. Finally, we note that our choice of parameters in Fig. 11(c) (SM [16] Sec. VII) was such that R_g (entering in the scaled sparsity) was significantly different from simply an average radius. Thus, the agreement with the theory illustrates the important role of the Gaussian geometry in the sparse case.

As discussed above, in the (high and moderate) sparse regimes, a large bias alters the anchor geometry of the two classes in different ways. To illustrate this important aspect, we show in Fig. 12 the effect of sparsity and bias on the
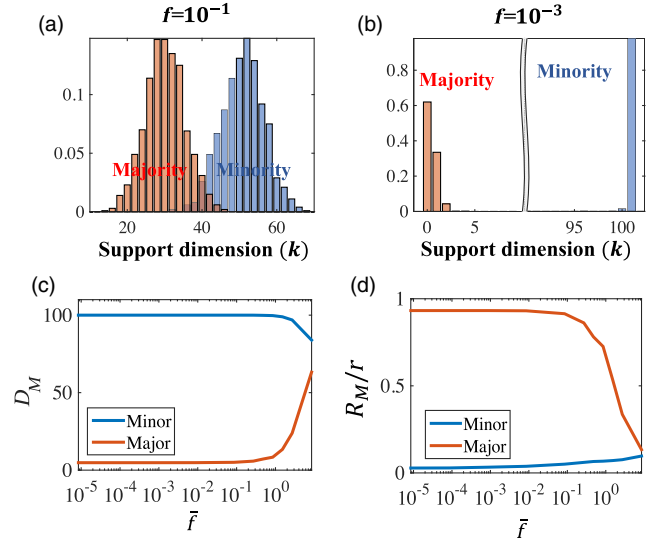
FIG. 12. Manifold configurations and geometries for classification of ℓ_1 ellipsoids with sparse labels, analyzed separately in terms of the majority and minority classes. The radii for ℓ_1 ellipsoids are $R_i = r$ for $1 \le i \le 10$ and $R_i = \frac{1}{2}r$ for $11 \le i \le 100$, with $r = 10$. (a) Histogram of support dimensions for moderate sparsity $f = 0.1$: minority (blue) and majority (red) manifolds. (b) Histogram of support dimensions for high sparsity $f = 10^{-3}$: minority (blue) and majority (red) manifolds. (c) Manifold dimension as a function of $\bar{f} = f(1 + R_g^2)$ as f is varied: minority (blue) and majority (red) manifolds. (d) Manifold radius relative to the scaling factor r as a function of $\bar{f}$: minority (blue) and majority (red) manifolds.

FIG. 11. (a),(b) Classification of ℓ_2 balls with sparse labels. (a) Capacity of ℓ_2 balls as a function of $\bar{f} = f(1 + r^2)$, for $D = 100$ (red) and $D = 10$ (blue) mean-field theory results. Overlaid black dotted lines represent an approximation interpolating between Eqs. (55) and (57) (details in SM [16], Sec. S9). (b), (c) Classification of general manifolds with sparse labels. (b) Capacity of ℓ_1 ellipsoids with $D = 100$, where the first 10 components are equal to r, and the remaining 90 components are $\frac{1}{2}r$, as a function of $\bar{f} = f(1 + R_g^2)$. r is varied from 10^{-3} to 10^3: numerical simulations (circles), mean-field theory (lines), spherical approximation (dotted lines). (c) Capacity of ring manifolds with a Gaussian falloff spectrum, with $\sigma = 0.1$ and $D = 100$ (details in SM [16]). The Fourier components in these manifolds have a Gaussian falloff, i.e., $R_n = A \exp\left(-\frac{1}{2}[2\pi(n-1)\sigma]^2\right)$.

geometry of the ℓ_1 ellipsoids studied in Fig. 11(c). Here, we show the evolution of R_M and D_M for the majority and minority classes as $\bar{f}$ increases. Note that, despite the fact that the shape of the manifolds is the same for all f, their *manifold anchor geometry* depends on both class membership and sparsity levels, because these measures depend on the margin. When $\bar{f}$ is small, the minority class has $D_M = D$, as seen in Fig. 12(c); i.e., the minority class manifolds are close to be fully supporting due to the large positive margin. This can also be seen in the distributions of support dimension shown in Figs. 12(a) and 12(b). On the other hand, the majority class has $D_M \approx D_g = 2 \log D_1 \ll D$, and these manifolds are mostly in the interior regime. As f increases, the geometrical statistics for the two classes become more similar. This is seen in Figs. 12(c) and 12(d), where D_M and R_M for both majority and minority classes converge to the zero margin value for large $\bar{f} = f(1 + R_g^2)$.

VII. SUMMARY AND DISCUSSION

Summary.— We have developed a statistical mechanical theory for linear classification of inputs organized in perceptual manifolds, where all points in a manifold share the same label. The notion of perceptual manifolds is critical in a variety of contexts in computational neuroscience modeling and in signal processing. Our theory is not restricted to manifolds with smooth or regular geometries; it applies to any compact subset of a D-dimensional affine subspace. Thus, the theory is applicable to manifolds arising from any variation in neuronal responses with a continuously varying physical variable, or from any sampled set arising from experimental measurements on a limited number of stimuli.

The theory describes the capacity of a linear classifier to separate a dichotomy of general manifolds (with a given margin) by a universal set of mean-field equations. These equations may be solved analytically for simple geometries, but, for more complex geometries, we have developed iterative algorithms to solve the self-consistent equations. The algorithms are efficient and converge fast, as they involve only solving for $O(D)$ variables of a single manifold, rather than invoking simulations of a full system of P manifolds embedded in $\mathbb{R}^N$.

Applications:—The statistical mechanical theory of perceptron learning has long provided a basis for understanding the performance and fundamental limitations of single-layer neural architectures and their kernel extensions. However, the previous theory only considered a finite number of random points, with no underlying geometric structure, and could not explain the performance of linear classifiers on a large, possibly infinite number of inputs organized as distinct manifolds by the variability due to changes in the physical parameters of objects. The theory presented in this work can explain the capacity and limitations of the linear classification of general manifolds.

This new theory is important for the understanding of how sensory neural systems perform invariant perceptual discrimination and recognition tasks of realistic stimuli. Furthermore, beyond estimating the classification capacity, our theory provides theoretically based geometric measures for assessing the quality of the neural representations of the perceptual manifolds. There are a variety of ways for defining the geometry of manifolds. Our geometric measures are unique in that they determine the ability to linearly separate the manifold, as our theory shows.

Our theory focuses on linear classification. However, it has broad implications for nonlinear systems and, in particular, for deep networks. First, most models of sensory discrimination and recognition in biological and artificial deep architectures model the readouts of the networks as linear classifiers operating on the top sensory layers. Thus, our manifold classification capacity and geometry can be applied to understand the performance of the deep network. Furthermore, the computational advantage of having multiple intermediate layers can be assessed by comparing the performance of a hypothetical linear classifier operating on these layers. In addition, the changes in the quality of representations across the deep layers can be assessed by comparing the changes in the manifold's geometries across layers. Indeed, previous discussions of sensory processing in deep networks hypothesized that neural object representations become increasingly untangled as the signal propagates along the sensory hierarchy. However, no concrete measure of untangling has been provided. The geometric measures derived from our theory can be used to quantify the degree of entanglement, tracking how the perceptual manifolds are nonlinearly reformatted as they propagate through the multiple layers of a neural network to eventually allow for linear classification in the top layer. Notably, the statistical, population-based nature of our geometric measures renders them ideally suited for comparison between layers of different sizes and nonlinearities, as well as between different deep networks or between artificial networks and biological ones. Lastly, our theory can suggest new algorithms for building deep networks, for instance, by imposing successive reduction of manifold dimensions and radii as part of an unsupervised learning strategy.

We have discussed above the domain of visual object classification and recognition, which has received immense attention in recent years. However, we would like to emphasize that our theory can be applied for modeling neural sensory processing tasks in other modalities. For instance, it can be used to provide insight on how the olfactory system performs discrimination and recognition of odor identity in the presence of orders-of-magnitude variations in odor concentrations. Neuronal responses to sensory signals are not static, but vary in time. Our theory can be applied to explain how the brain correctly decodes the stimulus identity despite its temporal nonstationarity.

Some of these applications may require further extensions of the present theory. The most important ones (currently the subject of ongoing work) include the following.

Correlations:— In the present work, we have assumed that the directions of the affine subspaces of the different manifolds are uncorrelated. In realistic situations, we expect to see correlations in the manifold geometries, mainly of two types. One is center-center correlations. Such correlations can be harmful for linear separability [32,33]. Another is correlated variability, in which the directions of the affine subspaces are correlated but not the centers. Positive correlations of the latter form are beneficial for separability. In the extreme case when the manifolds share a common affine subspace, the rank of the union of the subspaces is $D_{tot} = D$, rather than $D_{tot} = PD$, and the solution weight vector need only lie in the null space of this smaller subspace. Further work is needed to extend the present theory to incorporate more general correlations.

Generalization performance:— We have studied the separability of manifolds with known geometries. In many realistic problems, this information is not readily available and only samples reflecting the natural variability of input patterns are provided. These samples can be used to estimate the underlying manifold model (using manifold learning techniques [34,35]) and/or to train a classifier based upon a finite training set. Generalization error describes how well a classifier trained on a finite number of samples would perform on other test points drawn from the manifolds [36]. It would be important to extend our theory to calculate the expected generalization error achieved by the maximum margin solution trained on *point cloud manifolds*, as a function of the size of the training set and the geometry of the underlying full manifolds.

Unrealizable classification:—Throughout the present work, we have assumed that the manifolds are separable by a linear classifier. In realistic problems, the load may be above the capacity for linear separation, i.e., $\alpha > \alpha_M(\kappa = 0)$. Alternatively, neural noise may cause the manifolds to be unbounded in extent, with the tails of their distribution overlapping so that they are not separable with zero error. There are several ways to handle this issue in supervised learning problems. One possibility is to nonlinearly map the unrealizable inputs to a higher-dimensional feature space, via a multilayer network or nonlinear kernel function, where the classification can be performed with zero error. The design of multilayer networks could be facilitated using manifold processing principles uncovered by our theory.

Another possibility is to introduce an optimization problem allowing a small training error, for example, using a SVM with complementary slack variables [14]. These procedures raise interesting theoretical challenges, including understanding how the geometry of manifolds changes as they undergo nonlinear transformations, as well as investigating, by statistical mechanics, the performance of a linear classifier of manifolds with slack variables [37].

In conclusion, we believe the application of this theory and its corollary extensions will precipitate novel insights into how perceptual systems, biological or artificial, efficiently code and process complex sensory information.

APPENDIX A: REPLICA THEORY OF MANIFOLD CAPACITY

In this section, we outline the derivation of the mean-field replica theory summarized in Eqs. (11) and (12). We define the capacity of linear classification of manifolds, $\alpha_M(\kappa)$, as the maximal load, $\alpha = (P/N)$, for which with high probability a solution to $y^\mu \mathbf{w} \cdot \mathbf{x}^\mu \geq \kappa$ exists for a given κ. Here, $\mathbf{x}^\mu$ are points on the P manifolds M^μ, Eq. (1), and we assume that all $NP(D+1)$ components of $\{\mathbf{u}_i^\mu\}$ are drawn independently from a Gaussian distribution with zero mean and variance $(1/N)$, and that the binary labels $y^\mu = \pm 1$ are randomly assigned to each manifold with equal probabilities. We consider the thermodynamic limit, where $N, P \to \infty$, but $\alpha = (P/N)$, and D is finite.

Note that the geometric margin κ', defined as the distance from the solution hyperplane, is given by $y^\mu \mathbf{w} \cdot \mathbf{x}^\mu \geq \kappa' \|\mathbf{w}\| = \kappa' \sqrt{N}$. However, this distance depends on the scale of the input vectors $\mathbf{x}^\mu$. The correct scaling of the margin in the thermodynamic limit is $\kappa' = (\|\mathbf{x}\|/\sqrt{N})\kappa$. Since we adopted the normalization of $\|\mathbf{x}^\mu\| = O(1)$, the correct scaling of the margin is $y^\mu \mathbf{w} \cdot \mathbf{x}^\mu \geq \kappa$.

Evaluation of solution volume:—Following Gardner's replica framework, we first consider the volume Z of the solution space for $\alpha < \alpha_M(\kappa)$. We define the signed projections of the ith direction vector $\mathbf{u}_i^\mu$ on the solution weight as $H_i^\mu = \sqrt{N} y^\mu \mathbf{w} \cdot \mathbf{u}_i^\mu$, where $i = 1, \ldots, D+1$ and $\mu = 1, \ldots, P$. Then, the separability constraints can be written as $\sum_{i=1}^{D+1} S_i H_i^\mu \geq \kappa$. Hence, the volume can be written as

$$Z = \int d^N \mathbf{w} \, \delta(\mathbf{w}^2 - N) \Pi_{\mu=1}^P \Theta_\mu(g_S(\vec{H}^\mu) - \kappa), \quad \text{(A1)}$$

where $\Theta(x)$ is a Heaviside step function. g_S is the support function of S defined for Eq. (12) as $g_S(\vec{V}) = \min_{\vec{S}} \{\vec{V} \cdot \vec{S} | \vec{S} \in S\}$.

The volume defined above depends on the quenched random variables $\mathbf{u}_i^\mu$ and y^μ through H_i^μ. It is well known that, in order to obtain the typical behavior in the thermodynamic limit, we need to average $\log Z$, which we carry out using the replica trick, $\langle \log Z \rangle = \lim_{n\to 0} (\langle Z^n \rangle - 1)/n$, where $\langle \rangle$ refers to the average over $\mathbf{u}_i^\mu$ and y^μ. For natural n, we need to evaluate

$$\langle Z^n \rangle = \int \prod_\alpha^n d\mathbf{w}_\alpha \delta(\mathbf{w}_\alpha^2 - N) \prod_\mu^P \int \mathbb{D}\vec{H}^{\mu\alpha}$$

$$\times \left\langle \prod_i^{D+1} \sqrt{2\pi}\, \delta(H_i^{\mu\alpha} - y^\mu w_\alpha^T \mathbf{u}_i^\mu) \right\rangle_{\mathbf{u}_i^\mu, y^\mu}, \quad (A2)$$

where we have used the notation

$$\mathbb{D}\vec{H} = \Pi_{i=1}^{D+1} \frac{dH_i}{\sqrt{2\pi}} \Theta(g_S(\vec{H}) - \kappa). \quad (A3)$$

Using a Fourier representation of the delta functions, we obtain

$$\langle Z^n \rangle = \int \prod_\alpha^n d\mathbf{w}_\alpha \delta(\mathbf{w}_\alpha^2 - N) \prod_\mu^P \int \mathbb{D}\vec{H}^{\mu\alpha}$$

$$\times \prod_{i=1}^{D+1} \int \frac{d\hat{H}_i^{\mu\alpha}}{\sqrt{2\pi}} \langle \exp\{i\hat{H}_i^{\mu\alpha}(H_i^{\mu\alpha} - y^\mu \mathbf{w}_\alpha^T \mathbf{u}_i^\mu)\}\rangle_{\mathbf{u}_i^\mu, y^\mu}. \quad (A4)$$

Performing the average over the Gaussian distribution of $\mathbf{u}_i^\mu$ (each of the N components has zero mean and variance $(1/N)$) yields

$$\left\langle \exp \sum_{i=1}^{D+1} \sum_{\mu\alpha} \left[i\hat{H}_i^{\mu\alpha}\left(-y^\mu \sum_{j=1}^N w_\alpha^j \mathbf{u}_{i,j}^\mu\right)\right]\right\rangle_{\mathbf{u}_i^\mu, y^\mu}$$

$$= \exp\left\{-\frac{1}{2}\sum_{\alpha\beta} \mathfrak{q}_{\alpha\beta} \sum_{i\mu} \hat{H}_i^{\mu\alpha} \hat{H}_i^{\mu\beta}\right\}, \quad (A5)$$

where $\mathfrak{q}_{\alpha\beta} = (1/N)\sum_{j=1}^N w_\alpha^j w_\beta^j$. Thus, integrating the variables $\hat{H}_i^{\mu\alpha}$ yields

$$\langle Z^n \rangle = \int \prod_{\alpha=1}^n d\mathbf{w}_\alpha \delta(\mathbf{w}_\alpha^2 - N) \int d\mathfrak{q}_{\alpha\beta} \Pi_{\alpha\beta}$$

$$\cdot \delta(N\mathfrak{q}_{\alpha\beta} - \mathbf{w}_\alpha^T \mathbf{w}_\beta)\left[\exp\left(-\frac{(D+1)}{2}\log\det\mathfrak{q}\right)X\right]^P, \quad (A6)$$

where

$$X = \int \prod_\alpha \mathbb{D}\vec{H}^\alpha \exp\left[-\frac{1}{2}\sum_{i,\alpha,\beta} H_i^\alpha (\mathfrak{q}^{-1})_{\alpha\beta} H_i^\beta\right], \quad (A7)$$

and we have used the fact that all manifolds contribute the same factor.

We proceed by making the replica symmetric ansatz on the order parameter $\mathfrak{q}_{\alpha\beta}$ at its saddle point, $\mathfrak{q}_{\alpha\beta} = (1-q)\delta_{\alpha\beta} + q$, from which one obtains, in the $n \to 0$ limit,

$$\mathfrak{q}_{\alpha\beta}^{-1} = \frac{1}{1-q}\delta_{\alpha\beta} - \frac{q}{(1-q)^2} \quad (A8)$$

and

$$\log\det\mathfrak{q} = n\log(1-q) + \frac{nq}{1-q}. \quad (A9)$$

Thus, the exponential term in X can be written as

$$\exp\left[-\frac{1}{2}\sum_{\alpha i}\frac{(H_i^\alpha)^2}{1-q} + \frac{1}{2}\sum_i\left(\frac{\sqrt{q}}{1-q}\sum_\alpha H_i^\alpha\right)^2\right]. \quad (A10)$$

Using the Hubbard-Stratonovich transformation, we obtain

$$X = \int D\vec{T}\left[\int \mathbb{D}\vec{H}\exp\left\{-\frac{1}{2}\frac{\vec{H}^2}{1-q} + \frac{\sqrt{q}}{1-q}\vec{H}\cdot\vec{T}\right\}\right]^n, \quad (A11)$$

where $D\vec{T} = \Pi_i(dT_i/\sqrt{2\pi})\exp[-(T_i^2/2)]$. Completing the square in the exponential and using $\int D\vec{T}A^n = \exp n \int D\vec{T}\log A$ in the $n \to 0$ limit, we obtain $X = \exp\{[nq(D+1)/2(1-q)]\} + n\int D\vec{T}\log z(\vec{T}))$, with

$$z(\vec{T}) = \int \mathbb{D}\vec{H}\exp\left\{-\frac{1}{2(1-q)}||\vec{H} - \sqrt{q}\vec{T}||^2\right\}. \quad (A12)$$

Combining these terms, we write the last factor in Eq. (A6) as $\exp nPG_1$, where

$$G_1 = \int D\vec{T}\log z(\vec{T}) - \frac{(D+1)}{2}\log(1-q). \quad (A13)$$

The first factors in $\langle Z^n \rangle$, Eq. (A6), can be written as $\exp nNG_0$, whereas, in the Gardner theory, the entropic term in the thermodynamic limit is

$$G_0(q) = \frac{1}{2}\ln(1-q) + \frac{q}{2(1-q)} \quad (A14)$$

and represents the constraints on the volume of $\mathbf{w}_\alpha$ due to normalization and the order parameter $\mathfrak{q}$. Combining the G_0 and G_1 contributions, we have

$$\langle Z^n \rangle_{t_0,t} = e^{Nn[G_0(q) + \alpha G_1(q)]}. \quad (A15)$$

The classification constraints contribute αG_1, with Eq. (A13), and

$$z(\vec{T}) = \int \Pi_{i=1}^{D+1} \frac{dY_i}{\sqrt{2\pi(1-q)}}$$

$$\times \exp\left(-\frac{\vec{Y}^2}{2(1-q)}\right) \Theta(g_S(\sqrt{q}\vec{T} + \vec{Y}) - \kappa), \quad \text{(A16)}$$

where we have written the fields H_i as

$$H_i = \sqrt{q}T_i + Y_i. \quad \text{(A17)}$$

Note that $\sqrt{q}T_i$ represents the quenched random component due to the randomness in the $\mathbf{u}_i^\mu$, and Y_i is the "thermal" component due to the variability within the solution space. The order parameter q is calculated via $0 = (\partial G_0/\partial q) + \alpha(\partial G_1/\partial q)$.

Capacity:—In the limit where $\alpha \to \alpha_M(\kappa)$, the overlap between the solutions becomes unity and the volume shrinks to zero. It is convenient to define $Q = q/(1-q)$ and study the limit of $Q \to \infty$. In this limit, the leading order is

$$\langle \log Z \rangle = \frac{Q}{2}[1 - \alpha\langle F(\vec{T})\rangle_{\vec{T}}], \quad \text{(A18)}$$

where the first term is the contribution from $G_0 \to (Q/2)$. The second term comes from $G_1 \to -(Q/2)\alpha\langle F(\vec{T})\rangle_{\vec{T}}$, where the average is over the Gaussian distribution of the $(D+1)$-dimensional vector $\vec{T}$, and

$$F(\vec{T}) \to -\frac{2}{Q}\log z(\vec{T}) \quad \text{(A19)}$$

is independent of Q and is given by replacing the integrals in Eq. (A16) by their saddle point, which yields

$$F(\vec{T}) = \min_{\vec{V}}\{\|\vec{V} - \vec{T}\|^2 | g_S(\vec{V}) - \kappa \geq 0\}. \quad \text{(A20)}$$

At the capacity, $\log Z$ vanishes, and the capacity of a general manifold with margin κ is given by

$$\alpha_M^{-1}(\kappa) = \langle F(\vec{T})\rangle_{\vec{T}} \quad \text{(A21)}$$

$$F(\vec{T}) = \min_{\vec{V}}\{\|\vec{V} - \vec{T}\|^2 | g_S(\vec{V}) - \kappa \geq 0\}. \quad \text{(A22)}$$

Finally, we note that the mean squared "annealed" variability in the fields due to the entropy of solutions vanishes at the capacity limit, as $1/Q$; see Eq. (A16). Thus, the quantity $\|\vec{V} - \vec{T}\|^2$ in the above equation represents the annealed variability times Q, which remains finite in the limit of $Q \to \infty$.

APPENDIX B: STRICTLY CONVEX MANIFOLDS

1. General

Here, we evaluate the capacity of strictly convex manifolds, starting from the expression for general manifolds, Eq. (20). In a strictly convex manifold $\mathcal{S}$, any point in the line segment connecting two points $\vec{x}$ and $\vec{y}$, $\vec{x}, \vec{y} \in \mathcal{S}$, other than $\vec{x}$ and $\vec{y}$, belongs to the interior of $\mathcal{S}$. Thus, the boundary of the manifold does not contain edges or flats with spanning dimension $k > 1$. Indeed, the only possible spanning dimensions for the entire manifold are $k = 0, 1$, and $D + 1$. Therefore, there are exactly two contributions to the inverse capacity. When t_0 obeys $t_{\text{touch}}(\vec{t}) > t_0 - \kappa > t_{\text{fs}}(\vec{t})$, the integrand of Eq. (20) contributes $[(-\vec{t} \cdot \tilde{s}(\vec{T}) - t_0 + \kappa)^2/(1 + \|\tilde{s}(\vec{T})\|^2)]$. When $t_0 < \kappa + t_{\text{fs}}(\vec{t})$, the manifold is fully embedded. In this case, $\vec{v} = 0$, and the integrand reduces to Eq. (26). In summary, the capacity for convex manifolds can be written as

$$\alpha^{-1}(\kappa) = \int D\vec{t} \int_{\kappa+t_{\text{fs}}(\vec{t})}^{\kappa+t_{\text{touch}}(\vec{t})} Dt_0 \frac{(-\vec{t} \cdot \tilde{s}(\vec{t}, t_0) - t_0 + \kappa)^2}{1 + \|\tilde{s}(\vec{t}, t_0)\|^2}$$

$$+ \int D\vec{t} \int_{-\infty}^{\kappa+t_{\text{fs}}(\vec{t})} Dt_0[(t_0 - \kappa)^2 + \|\vec{t}\|^2], \quad \text{(B1)}$$

where $t_{\text{touch}}(\vec{t})$ and $t_{\text{fs}}(\vec{t})$ are given by Eqs. (22) and (25), respectively, and $\tilde{s}(\vec{t}, t_0) = \arg\min_{\vec{s}}(\vec{v} \cdot \vec{s})$, with $\vec{v} = \vec{t} + (v_0 - t_0)\tilde{s}$.

2. ℓ_2 balls

In the case of ℓ_2 balls with D and radius R, $g(\vec{v}) = -R\|\vec{v}\|$. Hence, $t_{\text{touch}}(\vec{t}) = R\|\vec{t}\|$ and $t_{\text{fs}}(\vec{t}) = -R^{-1}\|\vec{t}\|$. Thus, Eq. (B1) reduces to the capacity of balls is

$$\alpha_{\text{Ball}}^{-1} = \int_0^\infty dt\chi_D(t) \int_{\kappa-tR^{-1}}^{\kappa+tR} Dt_0 \frac{(-t_0 + tR + \kappa)^2}{(1 + R^2)}$$

$$+ \int_0^\infty dt\chi_D(t) \int_{-\infty}^{\kappa-tR^{-1}} Dt_0[(t_0 - \kappa)^2 + t^2], \quad \text{(B2)}$$

where

$$\chi_D(t) = \frac{2^{1-\frac{D}{2}}}{\Gamma(\frac{D}{2})} t^{D-1} e^{-\frac{1}{2}t^2}, \quad t \geq 0 \quad \text{(B3)}$$

is the D-dimensional Chi probability density function, reproducing the results of Ref. [15]. Furthermore, Eq. (B2) can be approximated by the capacity of points with a margin increase of $R\sqrt{D}$, i.e., $\alpha_{\text{Ball}}(\kappa, R_M, D_M) \approx (1 + R^2)\alpha_0(\kappa + R\sqrt{D})$ (details are given in Ref. [15]).

3. ℓ_2 ellipsoids

a. Anchor points and support regimes

With ellipsoids, Eq. (35), the support function in Eq. (15) can be computed explicitly as follows. For a vector

$\vec{V} = (\vec{v}, v_0)$, with nonzero $\vec{v}$, the support function $g(\vec{v})$ is minimized by a vector $\vec{s}$, which occurs on the boundary of the ellipsoid; i.e., it obeys the equality constraint $f(\vec{s}) = 0$ with $\rho(\vec{s}) = \sum_{i=1}^{D} (s_i/R_i)^2 - 1$, Eq. (35). To evaluate g, we differentiate $-\sum_{i=1}^{D} s_i \vec{v}_i + \rho f(\vec{s})$ with respect to s_i, where ρ is a Lagrange multiplier enforcing the constraint, yielding

$$\tilde{s}_i = \frac{v_i R_i^2}{g(\vec{v})} \tag{B4}$$

and

$$g(\vec{v}) = -\|\vec{v} \circ \vec{R}\|, \tag{B5}$$

where $\vec{R}$ is the D-dimensional vector of the ellipsoid principal radii, and $\circ$ refers to the pointwise product of vectors, $(\vec{v} \circ \vec{R})_i = v_i R_i$. For a given $(\vec{t}, t_0)$, the vector $(\vec{v}, v_0)$ is determined by Eq. (13), and the analytic solution above can be used to derive explicit expressions for $\tilde{s}(\vec{t}, t_0)$ in the different regimes as follows.

Interior regime:—In the interior regime $\lambda = 0$, $\vec{v} = \vec{t}$, resulting in zero contribution to the inverse capacity. The anchor point is given by the following boundary point on the ellipse, given by Eq. (B4) with $\vec{v} = \vec{t}$. This regime holds for t_0, obeying the inequality $t_0 - \kappa \geq t_{\text{touch}}(\vec{t})$, with Eq. (22), yielding

$$t_{\text{touch}}(\vec{t}) = \|\vec{t} \circ \vec{R}\|. \tag{B6}$$

Touching regime:—Here, the anchor point is given by Eq. (B4) where $\vec{v} = \vec{t} + \lambda \tilde{s}$. Substituting $t_i + \lambda \tilde{s}_i$ for v_i in the numerator of that equation, and $g(\vec{v}) = \kappa - v_0 = \kappa - t_0 - \lambda$, yields

$$\tilde{s}_{\text{touch},i}(\vec{t}, t_0) = -\frac{t_i R_i^2}{\lambda(1 + R_i^2) + (-t_0 + \kappa)}, \tag{B7}$$

where the parameter λ is determined by the ellipsoidal constraint,

$$1 = \sum_{i=1}^{D} \frac{t_i^2 R_i^2}{[\lambda(1 + R_i^2) + (-t_0 + \kappa)]^2}. \tag{B8}$$

In this regime, the contribution to the capacity is given by Eqs. (16) and (17), with $\tilde{s}$ in Eq. (B7).

The touching regime holds for $t_{\text{touch}} > t_0 - \kappa > t_{\text{fs}}$, where $\lambda \to \kappa - t_0$, $\vec{v}$ vanishes, and the anchor point is the point on the boundary of the ellipsoid antiparallel to $\vec{t}$. Substituting this value of λ in Eq. (B8) yields

$$t_{\text{fs}}(\vec{t}) = -\sqrt{\sum_i \left(\frac{t_i}{R_i}\right)^2}. \tag{B9}$$

Fully supporting regime:—When $t_0 - \kappa < t_{\text{fs}}$, we have $\vec{v} = 0$, $v_0 = \kappa$, and $\lambda = t_0 - \kappa$, implying that the center, as well as the entire ellipsoid, is fully supporting the max

margin solution. In this case, the anchor point is antiparallel to $\vec{t}$ at the interior point, Eq. (24), and its contribution to the capacity is as in Eq. (26).

APPENDIX C: LIMIT OF LARGE MANIFOLDS

In the large size limit, $\|\tilde{s}\| \to \infty$, linear separation of manifolds reduces to linear separation of P random D-dimensional affine subspaces. When separating subspaces, all of them must be fully embedded in the margin plane; otherwise, they would intersect it and violate the classification constraints. However, the way large-size manifolds approach this limit is subtle. To analyze this limit, we note that, when $\|\tilde{s}\|$ is large, $g(\vec{v}) > -\|\tilde{s}\|\|\vec{v}\|$ and, from the condition that $g(\vec{v}) \geq (-v_0 + \kappa)$, we have $\|\vec{v}\| \leq [(-v_0 + \kappa)/\|\tilde{s}\|]$; i.e., $\|\vec{v}\|$ is $O(\|\tilde{s}\|^{-1})$. A small $\|\vec{v}\|$ implies that the affine basis vectors, except the center direction, are all either exactly or almost orthogonal to the solution weight vector. Since $\lambda\tilde{s} = -\vec{t} + \vec{v} \approx -\vec{t}$, it follows that $\tilde{s}$ is almost antiparallel to the Gaussian vector $\vec{t}$, and hence, $D_M \to D$; see Eq. (29). To elucidate the manifold support structure, we note first that, by Eq. (22), $t_{\text{touch}} \propto -\|\vec{t}\|\|\tilde{s}\| \to -\infty$; hence, the fractional volume of the interior regime is negligible, and the statistics is dominated by the embedded regimes. In fact, the fully embedded transition is given by $t_{\text{embed}} \approx -\kappa$ [see Eq. (25)], so that the fractional volume of the fully embedded regime is $H(-\kappa) = \int_{-\kappa}^{\infty} Dt_0$, and its contribution to inverse capacity is, therefore, $\int_{-\kappa}^{\infty} Dt_0[\langle\|\vec{t}\|^2\rangle + (t_0 + \kappa)^2] = H(-\kappa)D + \alpha_0^{-1}(\kappa)$. The remaining summed probability of the touching and partially embedded regimes ($k \geq 1$) is, therefore, $H(\kappa)$. In these regimes, $\lambda^2\|\tilde{s}\|^2 \approx \lambda^2\|\tilde{s}\|^2 \approx \|\vec{t}\|^2$, so that this regime contributes a factor of $\int_{-\infty}^{-\kappa} Dt_0\langle\|\vec{t}\|^2\rangle = H(\kappa)D$. Combining these two contributions, we obtain, for large sizes, $\alpha_M^{-1} = D + \alpha_0^{-1}(\kappa)$, consistent with Eq. (8).

APPENDIX D: HIGH-DIMENSIONAL MANIFOLDS

1. High-dimensional ℓ_2 ball

Before we discuss general manifolds in high dimensions, we focus on the simple case of high-dimensional balls, the capacity of which is given by Eq. (B2). When $D \gg 1$, $\chi_D(t)$, Eq. (B3) is centered around $t = \sqrt{D}$. Substituting $\chi_D(t) \approx \delta(t - \sqrt{D})$ yields

$$\alpha_{\text{Ball}}^{-1}(\kappa, R, D) = \int_{\kappa - \frac{\sqrt{D}}{R}}^{\kappa + R\sqrt{D}} Dt_0 \frac{(R\sqrt{D} + \kappa - t)_0^2}{R^2 + 1}$$

$$+ \int_{-\infty}^{\kappa - \frac{\sqrt{D}}{R}} Dt_0([t_0 - \kappa]^2 + D). \tag{D1}$$

As long as $R \ll \sqrt{D}$, the second term in Eq. (D1) vanishes and yields

$$\alpha_{\text{Ball}}^{-1}(\kappa, R, D) = \int_{-\infty}^{\kappa + R\sqrt{D}} Dt_0 \frac{(R\sqrt{D} + \kappa - t_0)^2}{R^2 + 1}. \quad \text{(D2)}$$

Here, we note that the t_0 term in the numerator is significant only if $R = O(D^{-1/2})$ or smaller, in which case the denominator is just 1. When R is of order 1, the term t_0 in the integrand is negligible. Hence, in both cases, we can write

$$\alpha_{\text{Ball}}^{-1}(\kappa, R, D) \approx \alpha_0^{-1}\left(\frac{\kappa + R\sqrt{D}}{\sqrt{1 + R^2}}\right). \quad \text{(D3)}$$

In this form, the intuition beyond the factor $\sqrt{1 + R^2}$ is clear, stemming from the fact that the distance of a point from the margin plane scales with its norm; hence, the margin entering the capacity should factor out this norm.

As stated above, Eq. (D3) implies a finite capacity only in the scaling regime, where $R\sqrt{D} = O(1)$. If, on the other hand, $R\sqrt{D} \gg 1$, Eq. (D2) implies

$$\alpha_{\text{Ball}}^{-1} = \frac{R^2 D}{1 + R^2} \quad \text{(D4)}$$

[where we have used the asymptote $\alpha_0^{-1}(x) \to x^2$ for large x], reducing to $\alpha^{-1} = D$ for large R.

The analysis above also highlights the support structure of the balls at large D. As long as $R \ll \sqrt{D}$, the fraction of balls that lie fully in the margin plane is negligible, as implied by the fact that $t_{fs} \approx -\sqrt{D}/R \to -\infty$. The overall fraction of interior balls is $H(\kappa + R\sqrt{D})$, whereas the fraction that touch the margin planes is $1 - H(\kappa + R\sqrt{D})$. Despite the fact that there are no fully supporting balls, the touching balls are almost parallel to the margin planes if $R \gg 1$; hence, the capacity reaches its lower bound.

Finally, the large manifold limit discussed in Appendix C is realized for $R \ll \sqrt{D}$. Here, $t_{fs} = -\kappa$, and the system is either touching or fully supporting with probabilities $H(\kappa)$ and $H(-\kappa)$, respectively.

2. General manifolds

To analyze the limit of high-dimensional general manifolds, we utilize the self-averaging of terms in Eq. (20), involving sums of the D components of $\vec{t}$ and $\tilde{s}$, i.e.,

$$\vec{t} \cdot \tilde{s} \approx \langle ||\tilde{s}|| \rangle \langle \vec{t} \cdot \hat{s} \rangle \approx R_M \sqrt{D_M} = \kappa_M. \quad \text{(D5)}$$

Also, $t_{\text{touch}} \approx \vec{t} \cdot \tilde{s} \approx \kappa_M$; hence, we obtain for the capacity

$$\alpha_M^{-1}(\kappa) \approx \frac{\langle [\kappa_M + \kappa - t_0]_+^2 \rangle_{t_0}}{R_M^2 + 1}$$

$$\approx \alpha_{\text{Ball}}^{-1}(\kappa, R_M, D_M), \quad \text{(D6)}$$

where the average is with respect to the Gaussian t_0. Evaluating R_M and D_M involves calculations of the

self-consistent statistics of the anchor points. This calculation is simplified in high dimensions. In particular, λ, Eq. (17), reduces to

$$\lambda \approx \frac{\langle [\kappa_M + \kappa - t_0]_+ \rangle_{t_0}}{1 + R_M^2}. \quad \text{(D7)}$$

Hence, it is approximately constant, independent of $\vec{T}$.

In deriving the above approximations, we used self-averaging in summations involving the D intrinsic coordinates. The full dependence on the longitudinal Gaussian t_0 should remain. Thus, R_M and D_M should, in fact, be substituted by $R_M(t_0)$ and $D_M(t_0)$, denoting Eqs. (28) and (29), with averaging only over $\vec{t}$. This would yield a more complicated expressions than Eqs. (D6) and (D7).

The reason why we can replace them by average quantities is the following: The potential dependence of the anchor radius and dimension on t_0 is via $\lambda(t_0)$. However, inspecting Eq. (D7), we note two scenarios. In the first one, R_M is small and κ_M is of order 1. In this case, because the manifold radius is small, the contribution $\lambda \tilde{s}$ is small and can be neglected. This is the same argument why, in this case, the geometry can be replaced by the Gaussian geometry, which does not depend on t_0 or κ. The second scenario is that R_M is of order 1 and $\kappa_M \gg 1$, in which case the order 1 contribution from t_0 is negligible.

APPENDIX E: CAPACITY OF ℓ_2 BALLS WITH SPARSE LABELS

First, we note that the capacity of sparsely labeled points is $\alpha_0(f, \kappa) = \max_b \alpha_0(f, \kappa, b)$, where

$$\alpha_0^{-1}(f, \kappa, b) = f \int_{-\infty}^{\kappa + b} Dt(-t + \kappa + b)^2$$

$$+ (1 - f) \int_{-\infty}^{\kappa - b} Dt(-t + \kappa - b)^2. \quad \text{(E1)}$$

Optimizing b yields the following equation for b:

$$0 = f \int_{-\infty}^{b + \kappa} Dt(-t + \kappa + b)$$

$$+ (1 - f) \int_{-\infty}^{b - \kappa} Dt(-t + \kappa - b). \quad \text{(E2)}$$

In the limit of $f \to 0$, $b \gg 1$. The first equation reduces to

$$\alpha_0^{-1} \approx fb^2 + \exp(-b^2/2), \quad \text{(E3)}$$

yielding for the optimal $b \approx \sqrt{2|\log f|}$. With this b, the inverse capacity is dominated by the first term, which is $\alpha_0^{-1} \approx 2f|\log f|$.

The capacity for ℓ_2 balls of radius R, with sparse labels, with sparsity f [Eqs. (53) and (B2)], is given by

$$\alpha_{\text{Ball}}^{-1} = f \int_0^\infty dt \chi_D(t) \cdot \left[\int_{\kappa-tR^{-1}+b}^{\kappa+tR+b} Dt_0 \frac{(-t_0+tR+b+\kappa)^2}{(1+R^2)} \right.$$
$$+ \int_{-\infty}^{b+\kappa-tR^{-1}} Dt_0 [(t_0-b-\kappa)^2+t^2] \bigg]$$
$$+ (1-f) \int_0^\infty dt \chi_D(t) \cdot \left[\int_{\kappa-tR^{-1}-b}^{\kappa+tR-b} Dt_0 \frac{(-t_0+tR+\kappa-b)^2}{(1+R^2)} \right.$$
$$+ \int_{-\infty}^{-b+\kappa-tR^{-1}} Dt_0 [(t_0-\kappa+b)^2+t^2] \bigg], \qquad (E4)$$

and the optimal bias b is given by $\partial\alpha_{\text{Ball}}/\partial b = 0$. Here, we analyze these equations in various size regimes, assuming $\kappa = 0$.

Small R:—For balls with small radius, the capacity is that of points, unless the dimensionality is high. Thus, if D is large, but $R\sqrt{D} \lesssim 1$,

$$\alpha_{\text{Ball}}^{-1} = f \int_{-\infty}^{R\sqrt{D}+b} Dt_0 \frac{(-t_0+R\sqrt{D}+b)^2}{(1+R^2)}$$
$$+ (1-f) \int_{-\infty}^{R\sqrt{D}-b} Dt_0 \frac{(-t_0+R\sqrt{D}+b)^2}{(1+R^2)}; \quad (E5)$$

that is,

$$\alpha_{\text{Ball}}(f,R,D) \approx \alpha_0(f, \kappa = R\sqrt{D}). \qquad (E6)$$

As noted above, when $f \to 0$, the optimal bias diverges; hence the presence of order 1 in the induced margin is noticeable only for moderate f, such that $R\sqrt{D} = O(|\log f|)$.

Small f and large R:—Here, we analyze the above equations in the limit of small f and large R. We assume that f is sufficiently small so that the optimal bias b is large. The contribution of the minority class to the inverse capacity $\alpha_{\text{Ball}}^{-1}$ is dominated by

$$f \int_0^\infty dt \chi_D(t) \int_{-\infty}^b Dt_0 ((-t_0+b)^2+t^2) \approx fb^2. \quad (E7)$$

The dominant contribution of the majority class to $\alpha_{\text{Ball}}^{-1}$ is

$$(1-f) \int_0^\infty dt \chi_D(t) \int_{-b-Rt^{-1}}^{-b+tR} Dt_0 \frac{(-t_0+tR-b)^2}{(1+R^2)} \qquad (E8)$$

$$\approx (1-f) \int_{\bar{b}}^\infty dt \chi_D(t)(t-\bar{b})^2, \qquad (E9)$$

where $\bar{b} = b/R$. In deriving the last equation, we used $[(-t_0+tR-b)^2/(1+R^2)] \to (t-\bar{b})^2$ as $R, b \to \infty$. Second, the integrals are of substantial value only if $t \geq \bar{b}$, in which

case the integral of t_0 is $\int_{-R(\bar{b}-t)}^b Dt_0 \approx 1$, and the integral over t is from $\bar{b}$ to ∞. Combining the two results yields the following simple expression for the inverse capacity:

$$\alpha_{\text{Ball}}^{-1} = \bar{f}\bar{b}^2 + \int_{\bar{b}}^\infty dt \chi_D(t)(t-\bar{b})^2, \qquad (E10)$$

where scaled sparsity is $\bar{f} = fR^2$. The optimal scaled bias is given by

$$\bar{f}\bar{b} = \int_{\bar{b}}^\infty dt \chi_D(t)(t-\bar{b}). \qquad (E11)$$

Note that R and f affect the capacity only through the scaled sparsity. When $\bar{f} \to 0$, capacity is proportional to $(\bar{f}|\log\bar{f}|)^{-1}$. In a realistic regime of small $\bar{f}$ (between 10^{-4} and 1), capacity decreases with $\bar{f}$ roughly as $1/\bar{f}$, with a proportionality constant that depends on D, as shown in the examples of Fig. 12(a). Finally, when R is sufficiently large that $\bar{f} > 1$, $\bar{b}$ is order 1 or smaller. In this case, the second term in Eq. (E10) dominates and contributes $\langle t^2 \rangle = D$, yielding a capacity that saturates as D^{-1}, as shown in Fig. 11(a). It should be noted, however, that when b is not large, the approximations in Eqs. (E10) and (E11) no longer hold.

[1] J. J. DiCarlo and D. D. Cox, *Untangling Invariant Object Recognition*, Trends Cognit. Sci. **11**, 333 (2007).

[2] J. K. Bizley and Y. E. Cohen, *The What, Where and How of Auditory-Object Perception*, Nat. Rev. Neurosci. **14**, 693 (2013).

[3] K. A. Bolding and K. M. Franks, *Complementary Codes for Odor Identity and Intensity in Olfactory Cortex*, eLife **6**, e22630 (2017).

[4] H. S. Seung and D. D. Lee, *The Manifold Ways of Perception*, Science **290**, 2268 (2000).

[5] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, *How Does the Brain Solve Visual Object Recognition?*, Neuron **73**, 415 (2012).

[6] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, *Exponential Expressivity in Deep Neural Networks Through Transient Chaos*, in *Advances in Neural Information Processing Systems* (Morgan Kaufmann, San Francisco, CA, 2016), pp. 3360–3368.

[7] M. A. Ranzato, F. J. Huang, Y.-L. Boureau, and Y. LeCun, *Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition*, in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007), pp. 1–8.

[8] Y. Bengio, *Learning Deep Architectures for AI*, Foundations and Trends in Machine Learning **2**, 1 (2009).

[9] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng, *Measuring Invariances in Deep Networks*, in *Advances in Neural Information Processing Systems* (Morgan Kaufmann, San Francisco, CA, 2009), pp. 646–654.

[10] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, *Deep Neural*

Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition, PLoS Comput. Biol. **10**, e1003963 (2014).

[11] T. M. Cover, *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition*, IEEE Trans. Electron. Comput. **EC-14**, 326 (1965).

[12] E. Gardner, *The Space of Interactions in Neural Network Models*, J. Phys. A **21**, 257 (1988).

[13] E. Gardner, *Maximum Storage Capacity in Neural Networks*, Europhys. Lett. **4**, 481 (1987).

[14] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[15] S. Chung, D. D. Lee, and H. Sompolinsky, *Linear Readout of Object Manifolds*, Phys. Rev. E **93**, 060301 (2016).

[16] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevX.8.031003 for the further details of theory, and the details of numerics such as algorithms and the simulation parameters.

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

[18] F. Gerl and U. Krey, *Storage Capacity and Optimal Learning of Potts-Model Perceptrons by a Cavity Method*, J. Phys. A **27**, 7353 (1994).

[19] W. Kinzel and M. Opper, *Dynamics of Learning*, in *Models of Neural Networks* (Springer, 1991), pp. 149–171.

[20] R. T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, NJ, 2015).

[21] J.-J. Moreau, *Décomposition Orthogonale D'un Espace Hilbertien Selon Deux Cônes Mutuellement Polaires*, C.R. Hebd. Seances Acad. Sci. **225**, 238 (1962).

[22] S. Chung, U. Cohen, H. Sompolinsky, and D. D. Lee, *Learning Data Manifolds with a Cutting Plane Method*, Neural Comput., DOI: 10.1162/neco_a_01119 (2018).

[23] B. Grünbaum and G. C. Shephard, *Convex Polytopes*, Bull. Lond. Math. Soc. **1**, 257 (1969).

[24] A. A. Giannopoulos, V. D. Milman, and M. Rudelson, *Convex Bodies with Minimal Mean Width*, in *Geometric Aspects of Functional Analysis* (Springer, New York, 2000), pp. 81–93.

[25] R. Vershynin, *Estimation in High Dimensions: A Geometric Perspective*, in *Sampling Theory, A Renaissance* (Springer, New York, 2015), pp. 3–66.

[26] K. Rajan, L. F. Abbott, and H. Sompolinsky, *Stimulus-Dependent Suppression of Chaos in Recurrent Neural Networks*, Phys. Rev. E **82**, 011903 (2010).

[27] A. Litwin-Kumar, K. D. Harris, R. Axel, H. Sompolinsky, and L. F. Abbott, *Optimal Degrees of Synaptic Connectivity*, Neuron **93**, 1153 (2017).

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A Large-Scale Hierarchical Image Database*, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (IEEE, 2009), pp. 248–255.

[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going Deeper with Convolutions*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE Computer Society, 2015), pp. 1–9.

[30] S.-H. Kye, *Faces for Two Qubit Separable States and the Convex Hulls of Trigonometric Moment Curves*, arXiv:1302.2226.

[31] Z. Smilansky, *Convex Hulls of Generalized Moment Curves*, Isr. J. Math. **52**, 115 (1985).

[32] R. Monasson, *Properties of Neural Networks Storing Spatially Correlated Patterns*, J. Phys. A **25**, 3701 (1992).

[33] B. Lopez, M. Schroder, and M. Opper, *Storage of Correlated Patterns in a Perceptron*, J. Phys. A **28**, L447 (1995).

[34] S. T. Roweis and L. K. Saul, *Nonlinear Dimensionality Reduction by Locally Linear Embedding*, Science **290**, 2323 (2000).

[35] J. B. Tenenbaum, V. De Silva, and J. C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, Science **290**, 2319 (2000).

[36] S.-i. Amari, N. Fujita, and S. Shinomoto, *Four Types of Learning Curves*, Neural Comput. **4**, 605 (1992).

[37] S. Risau-Gusman and M. B. Gordon, *Statistical Mechanics of Learning with Soft Margin Classifiers*, Phys. Rev. E **64**, 031907 (2001).