Soft-margin classification of object manifolds

Uri Cohen [1] and Haim Sompolinsky [1,2,*]

[1] *Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem 9190401, Israel*

[2] *Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*

(Received 16 March 2022; accepted 29 July 2022; published 24 August 2022)

A neural population responding to multiple appearances of a single object defines a manifold in the neural response space. The ability to classify such manifolds is of interest, as object recognition and other computational tasks require a response that is insensitive to variability within a manifold. Linear classification of object manifolds was previously studied for max-margin classifiers. Soft-margin classifiers are a larger class of algorithms and provide an additional regularization parameter used in applications to optimize performance outside the training set by balancing between making fewer training errors and learning more robust classifiers. Here we develop a mean-field theory describing the behavior of soft-margin classifiers applied to object manifolds. Analyzing manifolds with increasing complexity, from points through spheres to general manifolds, a mean-field theory describes the expected value of the linear classifier's norm, as well as the distribution of fields and slack variables. By analyzing the robustness of the learned classification to noise, we can predict the probability of classification errors and their dependence on regularization, demonstrating a finite optimal choice. The theory describes a previously unknown phase transition, corresponding to the disappearance of a nontrivial solution, thus providing a soft version of the well-known classification capacity of max-margin classifiers. Furthermore, for high-dimensional manifolds of any shape, the theory prescribes how to define manifold radius and dimension, two measurable geometric quantities that capture the aspects of manifold shape relevant to soft classification.

DOI: 10.1103/PhysRevE.106.024126

I. INTRODUCTION

Max-margin and soft-margin classification When performing linear classification, the naive approach would aim for classifying all the training samples correctly with the largest possible margin, an approach known as max-margin classification [1,2]. An alternative approach, known as soft-margin classification [3,4], is to allow for misclassification of some of the samples, in order to increase the classification margin of most samples. Soft-margin classification is common in applications, where the data are not necessarily linearly separable. Furthermore, it allows for minimizing generalization error by optimizing a regularization parameter that balances between classification errors on the training set and achieving a larger margin. Both max-margin and soft-margin classification problems are solved by Support Vector Machine algorithms (SVMs).

Previous works on manifold classification The problem of manifold classification arises in neuroscience and machine learning when a population of biological or artificial neurons represents an object, and variability in object appearance would define a manifold in the neural response space. In invariant object recognition tasks, the response of output neurons is determined by object identity alone, which is naturally defined as performing manifold classification, i.e., using target labels that are constant within manifolds. The ability to perform max-margin classification on manifolds of increased complexity was analyzed in recent years. Building on the seminal work of Gardner [5] which considered the classification of points, recent works have extended theory to describe manifolds of any shape [6,7] and to allow for certain correlations between manifolds [8]. Those theoretical advances described only max-margin classifiers, which are not common in applications. Here we close this gap by analyzing soft classification of manifolds of increasing complexity, going from points, through spheres, to general manifolds.

Previous works on soft classification theory Previous theoretical works on soft-margin classifiers have analyzed the classification of finite numbers of training samples. Statistical learning tools were used to provide bounds on the generalization error and its asymptotic convergence toward the error of the Bayes optimal classifier [9–11]. Such analysis is used to compare different kernels and different regularization schemes [9], and an analysis of the behavior of the error shows how the choice of regularization can be used for improving upon the bounds available for max-margin classifiers [10]. A statistical physics analysis of soft-margin classification in a teacher-student setup described the learning curve, i.e., the dependence of training and generalization error on the number of samples (extending the max-margin analysis [12]). Such analysis was done for the unrealizable case where the teacher is more sophisticated than the student [13] and for realizable cases with or without noise [14]. Here we avoid making specific assumptions on the teacher and instead consider soft classification performance when averaging over random

*Corresponding author: haim@fiz.huji.ac.il

choice of labels. A different statistical physics approach analyzed the asymptotic behavior of max-margin and soft-margin classifiers [15]. It predicts classification error rates, assuming a large number of high-dimensional samples are drawn from a Gaussian mixture distribution.

The role of noise. When a soft-margin classifier is learned on a training set and then evaluated on a held-out test set, the classification errors achieved are called the training error and the test error, respectively. In general we expect the training error to be minimized for the max-margin classifier while the test error may be minimized at a finite value of the soft classification regularization parameter, which needs to be found empirically. Here we aim to analyze this setting by considering a test set that is a noisy version of the training set. This corresponds to noise resistance of the classifier, and not to the notion of generalization error in machine learning where it is assumed that the training and test set are sampled from the same distribution.

II. RESULTS

A. Soft classification of points

Max-margin classification of points is discussed by [5]; here we extend this seminal work to soft classification. Given P pairs $\{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^P$ of points $\boldsymbol{x}^\mu \in \mathbb{R}^N$ and labels $y^\mu \in \{\pm 1\}$, soft classification is defined by a set of weights $\boldsymbol{w} \in \mathbb{R}^N$ and slack variables $\vec{s} \in \mathbb{R}^P$ such that the fields at the solution obey for all $\mu \in [1, \ldots, P]$:

$$h^\mu = y^\mu \boldsymbol{w} \cdot \boldsymbol{x}^\mu \geqslant 1 - s^\mu. \tag{1}$$

The bold notation for $\boldsymbol{x}^\mu$ and $\boldsymbol{w}$ indicates that they are vectors in $\mathbb{R}^N$, whereas the arrow notation is used for other vectors, such as $\vec{s}$. Given a regularization parameter $c \geqslant 0$ the optimal classifier and slack variables are defined $\boldsymbol{w}^*, \vec{s}^* = \arg\min_{\boldsymbol{w},\vec{s}} L(\boldsymbol{w}, \vec{s})$ for a Lagrangian

$$L = \|\boldsymbol{w}\|^2/N + c\|\vec{s}\|^2/N \text{ s.t. } \forall \mu\ h^\mu \geqslant 1 - s^\mu, \tag{2}$$

and L^* denotes the minimal value of L.

Replica theory From the Lagrangian the volume of solutions $V(L, c)$ for a given value of the loss L and a choice of regularization c is given by

$$V(L, c) = \int d^N \boldsymbol{w} \int d^P \vec{s}\, \delta(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL) \tag{3}$$

$$\cdots \prod_\mu^P \delta(y^\mu \boldsymbol{w} \cdot \boldsymbol{x}^\mu - h^\mu)\Theta(h^\mu - 1 + s^\mu). \tag{4}$$

The volume is defined for any positive L, c, but we are interested in the problem parameters where it vanishes, which is expected to happen only at the minimal value L^*. Thus by analyzing the conditions where $V \to 0$ we characterize the optimal solution achieved by the optimization procedure, without introducing an additional temperature variable as is usually done (e.g., [12,16]). This allows us to describe not only L^* but also the expected norms of the weights $\|\boldsymbol{w}\|$ and slack variables $\|\vec{s}\|$, and the relation between N and P where the solution is achieved. For random labels $\vec{y} \in \{\pm 1\}^P$ and points $\boldsymbol{x}_i^\mu \sim \mathcal{N}(0, 1/N)$ we calculate the volume through replica identity:

$$[\log V]_{x,y} = \lim_{n\to 0} \left[\frac{V^n - 1}{n}\right]_{x,y}. \tag{5}$$

We solve this problem using a (replica symmetric) mean-field theory, which is expected to be exact in the thermodynamic limit $N, P \to \infty$ with a finite ratio $\alpha = P/N$. Replica symmetry is to be expected for soft classification since the problem is convex in both $\boldsymbol{w}$ and $\vec{s}$; hence the landscape does not have local minima and the global minima are either unique or form a convex set. In contrast, replica symmetry breaking implies the existence of multiple unconnected minima [17]. Analyzing the case where $V \to 0$ we obtain an expression for the loss L in terms of two order parameters q and k (see Appendix A 2; all notes are found in the Appendix):

$$L/q = \frac{k-1}{k} + \frac{c}{1+ck}\alpha\alpha_0^{-1}(1/\sqrt{q}), \tag{6}$$

where $\alpha_0^{-1}(\kappa) = \int_{-\infty}^{\kappa} Dt(\kappa - t)^2$ is Gardner's points capacity [5], $q = \|\boldsymbol{w}\|^2/N$ is the norm of the weight vector, and k is an additional order parameter, discussed below. Note we assumed here $\|\boldsymbol{x}^\mu\| = 1$; if instead $\|\boldsymbol{x}^\mu\| = a$, then q, c need to be scaled by $1/a^2$.

Self-consistent equations We expect the solution to satisfy saddle-point conditions $0 = \frac{\partial L}{\partial q} = \frac{\partial L}{\partial k}$, yielding two self-consistent equations for the order parameters q, k (see Appendix A 3):

$$1 = \frac{(ck)^2}{(1+ck)^2}\alpha\alpha_0^{-1}(1/\sqrt{q}), \tag{7}$$

$$1 - k = \frac{ck}{(1+ck)}\alpha H(-1/\sqrt{q}) \tag{8}$$

for $H(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} e^{-x^2/2}$ the Gaussian tail function.

The mean-field equations can be solved numerically for any load α (Algorithm 1; all algorithms are found in the Supplemental Material [18]); Figs. 1(a) and 1(b) show the resulting values of q and k, respectively. We observe that $k(\alpha)$ decreases monotonically from 1 at 0 [Fig. 1(b)], and similarly $\|\vec{s}\|$ increases monotonically from 0 to 1 [Fig. 1(c)]. Those are tightly related as from Eqs. (2), (6), and (7) we have that ck describes the relative strength of the weights' norm and the slack norm at the optimization target [Eq. (2)]:

$$ck = \sqrt{q/\alpha\langle s^2\rangle}. \tag{9}$$

Thus the loss is dominated by the weights when ck is large and by the slacks when it is small.

In contrast, $q(\alpha)$ is nonmonotonic, increasing from 0 to a peak at a finite value, then decreasing [Fig. 1(a)]. This is an indication of the trade-off between achieving a larger margin (small q) and making only small errors (small $\|\vec{s}\|$). Figure S1 (all figures numbered with an S are found in the Supplemental Material [18]) compares simulation results for q with solutions of the self-consistent equations.

We now consider some interesting limits (see Appendix A 3). When $\alpha \to 0$ we have $k \to 1$ and $q \to 0$ so that $\alpha_0^{-1}(1/\sqrt{q}) \approx 1/q$ and thus $k \approx 1 - \alpha c/(1+c)$ and $q \approx$

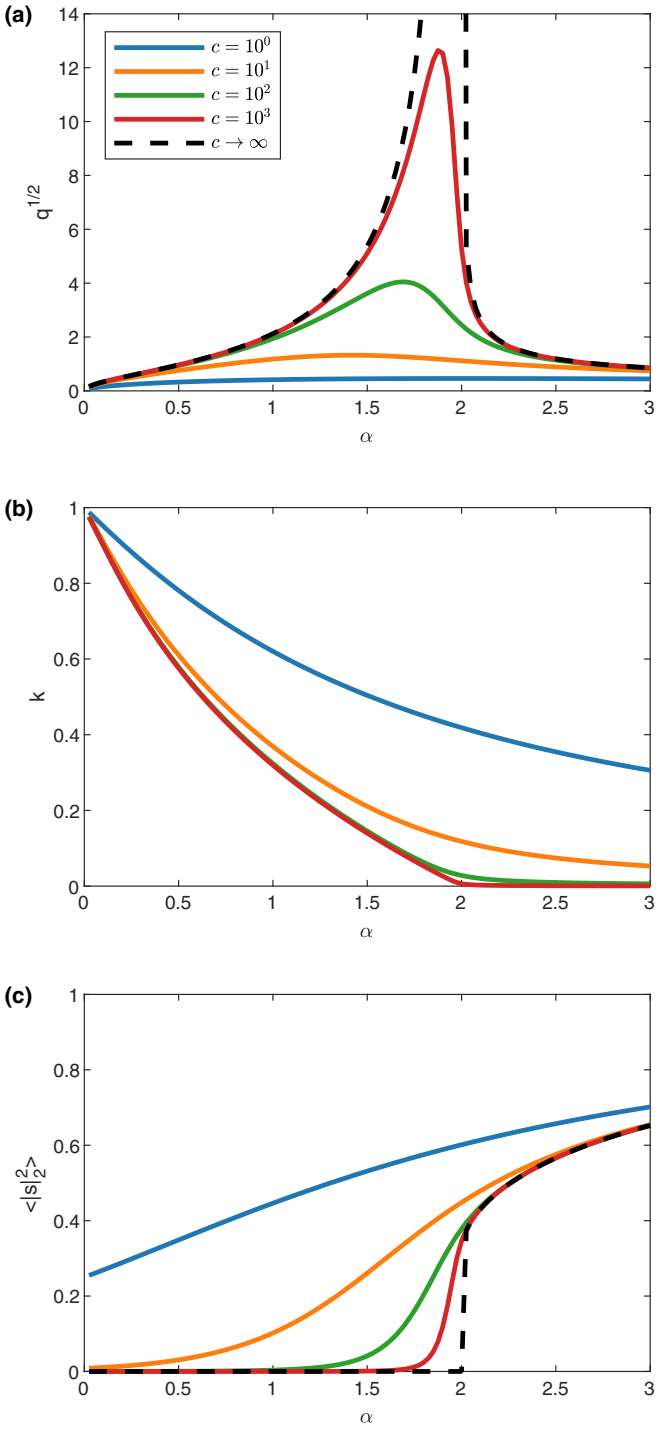

FIG. 1. Order parameters in soft classification of points. (a) The optimal weights' norm $q^{1/2}$ (y axis) for different values of α (x axis) and choices of the regularization variable c (color coded), including the $c \to \infty$ limit (dashed line). (b) The order parameter k (y axis) for different values of α (x axis) and choices of c (color coded). (c) The mean slack norm $\|\vec{s}\|^2$ (y axis) for different values of α (x axis) and choices of c (color coded), including the $c \to \infty$ limit (dashed line).

$\alpha c^2/(1+c)^2$. When $\alpha \to \infty$ we have $k \to 0$ and $q \to 0$ with scaling $k \approx 1/c\alpha$, $q \approx 1/\alpha$. Both limits are marked in Fig. S1.

Infinite c limit When $c \to \infty$ and $\alpha < 2$ there is a solution for $\boldsymbol{w}$ (of unconstrained norm) where $\vec{s} = \vec{0}$, so the Lagrangian becomes that of max-margin classifiers:

$$L = \min \|\boldsymbol{w}\|^2 \text{ s.t. } \forall \mu \; h^\mu \geqslant 1. \tag{10}$$

In this regime k is finite while ck diverges, so Eq. (7) recovers the max-margin theory [5] and q diverges for α near 2. On the other hand, for $c \to \infty$ and $\alpha > 2$ there is no solution with $\vec{s} = \vec{0}$ so this term dominates the loss and the Lagrangian becomes

$$L = \min \|\vec{s}\|^2 \text{ s.t. } \forall \mu \; h^\mu \geqslant 1 - s^\mu. \tag{11}$$

A mean-field solution of this Lagrangian involves two order parameters $q = \|\boldsymbol{w}\|^2/N$ and $K = \lim_{c\to\infty} ck$, which follow the self-consistent Eqs. (7) and (8) [where k on the left-hand-side of Eq. (8) approaches 0; see Appendix A 3]. Thus in the limit of $c \to \infty$ the mean-field theory reduces to a simple relation between q and α [dashed line in Fig. 1(a)]:

$$\alpha = \begin{cases} \alpha_0(1/\sqrt{q}) & \alpha < 2 \\ \alpha_0^{-1}(1/\sqrt{q})/H^2(-1/\sqrt{q}) & \alpha > 2 \end{cases}. \tag{12}$$

Field distribution The theory also provides the joint distribution of h, s; their variance is due to the quenched variability in the choice of the classification labels and the arrangement of points (see Appendix A 4). The field distribution is a concatenation of truncated Gaussian variables, each representing a different solution regime:

$$h \sim \begin{cases} \mathcal{N}\left(\frac{ck}{1+ck}, \frac{q}{(1+ck)^2}\right) & h < 1 \\ \mathcal{N}(0, q) & h \geqslant 1 \end{cases}. \tag{13}$$

Fields $h \geqslant 1$ are the "interior" regime (i.e., of points beyond the separating hyperplane), where $s = 0$, while fields $h < 1$ are the "touching" regime (i.e., of points touching the separating hyperplane), where $s > 0$. This distribution is shown for several choices of c and α in Fig. 2(a), and Fig. S2 compares theory to the empirical histogram from simulations. The distribution of slack variables then follows from $s = \max\{1 - h, 0\}$.

Classification errors We now turn our focus to the classification errors achieved when performing soft classification. The classification error on the training set is defined $\varepsilon_{tr} = P(h < 0) = P(s > 1)$, and from the field distribution we have

$$\varepsilon_{tr} = H(ck/\sqrt{q}) = H(1/\sqrt{\alpha \langle s^2 \rangle}). \tag{14}$$

A comparison of the training error observed in simulations with the theoretical predictions is given in Fig. S3. As demonstrated in Figs. 2(b) and 2(d), the training error is monotonically increasing with α and monotonically decreasing with c throughout. For $\alpha < 2$ where max-margin classifiers achieve no errors this is to be expected, but surprisingly this is also the case for $\alpha \geqslant 2$ [see classification error for $\alpha \geqslant 2$, $c \to \infty$ in Fig. 2(b)].

Thus we turn to analyze classification error in the presence of noise, where a finite c may be optimal. When Gaussian noise $\mathcal{N}(0, \sigma^2/N)$ is applied at each component of the input vectors, a noise $\mathcal{N}(0, \sigma^2 q)$ is added to the fields, so test error with respect to such noise is give by $\varepsilon_g = P(h + \eta\sigma\sqrt{q} < 0)$, where η is a standard Gaussian, or equivalently:

$$\varepsilon_g = \langle H(h/\sigma\sqrt{q}) \rangle_h. \tag{15}$$

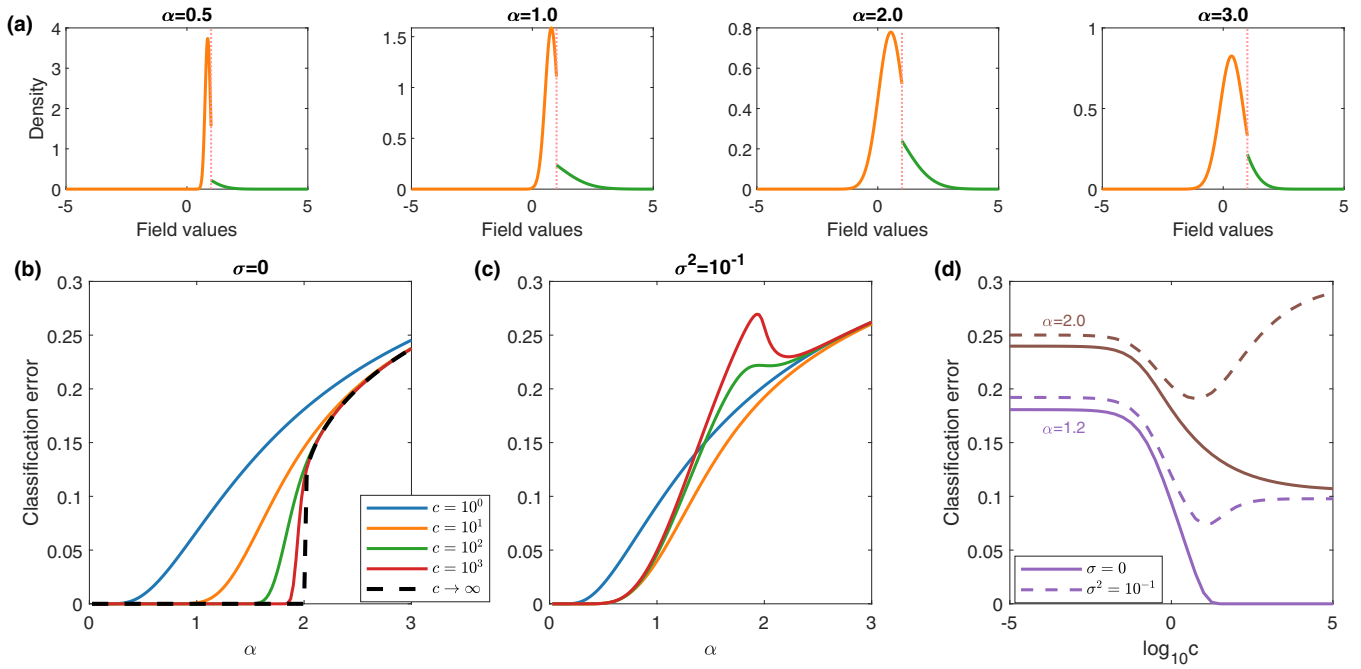

FIG. 2. Field distribution and errors in soft classification of points. (a) Field distributions at different values of α (panels), with color coded regime (orange: "touching" regime; green: "interior" regime; a dashed line at $h = 1$ indicates regime boundary), using $c = 10$. (b–c) Classification error (y axis) for different values of α (x axis) and choices of c (color coded), including the $c \to \infty$ limit [dashed line in (b)]. Each panel (b)–(c) shows the error at a different noise level σ^2 (indicated in title). (d) Classification error (y axis) for different choices of c (x axis, log scale), for several values of α (color coded), levels of noise σ^2 (solid/dashed lines).

Equation (15) can be evaluated using the field distribution [Eq. (13)]. The resulting levels of error exhibit nonmonotonic dependence on both α and c [Figs. 2(c) and 2(d)]. A comparison of the theoretical predictions with simulation results for different choices of c and levels of noise is provided in Fig. S4.

Classification errors for small noise While an explicit expression for the error is complicated, when the noise is small relative to the margin from the optimal hyperplane $\sigma \ll 1/\sqrt{q}$, we provide a simple approximation for the test error, which can be written as a signal-to-noise ratio $\varepsilon_g \approx H(\mathcal{S})$ (SNR; see Appendix A 5):

$$\mathcal{S} = ck/\sqrt{q[1 + (1 + ck)^2\sigma^2]}. \tag{16}$$

From the scaling of q, k for large and small α's we have

$$\mathcal{S} \approx \begin{cases} 1/\sqrt{\alpha[1/(1+c)^2 + \sigma^2]} & \alpha \ll 1 \\ 1/\sqrt{\alpha(1+\sigma^2)} & \alpha \gg 1 \end{cases}. \tag{17}$$

In this regime the optimal choice of c can be found by maximizing $\mathcal{S}$ [Eq. (16)] with respect to c, that is solving $0 = \frac{\partial \mathcal{S}^{-2}}{\partial c}$ for c, which yields (see Appendix A 6)

$$c^* = \frac{\sigma^{-2}}{1-k} - \frac{1}{k}, \tag{18}$$

which is positive in the regime where the SNR is a valid approximation, and needs to be solved self-consistently as k depends on c. Due to the dependence on k we have that c^* depends on α, but this analysis also suggests a "canonical choice" of c which is independent of α:

$$c \approx \sigma^{-2}. \tag{19}$$

This choice is expected to capture the order of magnitude of c^*, except when α is very small or very large [as Eq. (18) diverges for both $k \to 0$ and $k \to 1$].

Figure 3(a) demonstrates the optimal choice of c calculated by solving Eq. (18) and compares it to Eq. (19), showing this approximation is within the correct scale for a large range of α values. The resulting norm of the optimal solution changes smoothly with α [Fig. 3(b)], and the canonical choice of c achieves classification error which differs from the optimal one only when the error is much smaller than 1 [Fig. 3(c)] and is superior to other suboptimal choices of c (Fig. S5).

B. Methods for soft classification of manifolds

A manifold $M^\mu \subseteq \mathbb{R}^N$ for index $\mu \in [1, \ldots, P]$ is parameterized by its axes $\{\boldsymbol{u}_l^\mu \in \mathbb{R}^N\}_{l=0,\ldots,D}^{\mu=1,\ldots,P}$ and the manifold's intrinsic coordinates $\vec{S} \in \mathcal{M}^\mu \subseteq \mathbb{R}^{D+1}$. Each point in the manifold is a vector $\boldsymbol{x}^\mu(\vec{S}) \in M^\mu$ such that

$$\boldsymbol{x}^\mu(\vec{S}) = \sum_{l=0}^{D} \boldsymbol{u}_l^\mu S_l. \tag{20}$$

As above, the bold notation for $\boldsymbol{x}^\mu$ and $\boldsymbol{u}_l^\mu$ indicates that they are vectors in $\mathbb{R}^N$, whereas the arrow notation is used for other vectors, such as the coordinates $\vec{S}$ (not to be confused with the slack $\vec{s}$). By convention $\boldsymbol{u}_0^\mu$ is the manifold center and we take $S_0 = 1$, so that distances are measured in units of the center norm. When classifying P manifolds with weights $\boldsymbol{w} \in \mathbb{R}^N$, denoting axes projections $v_l^\mu = y^\mu \boldsymbol{u}_l^\mu \cdot \boldsymbol{w}$ the fields

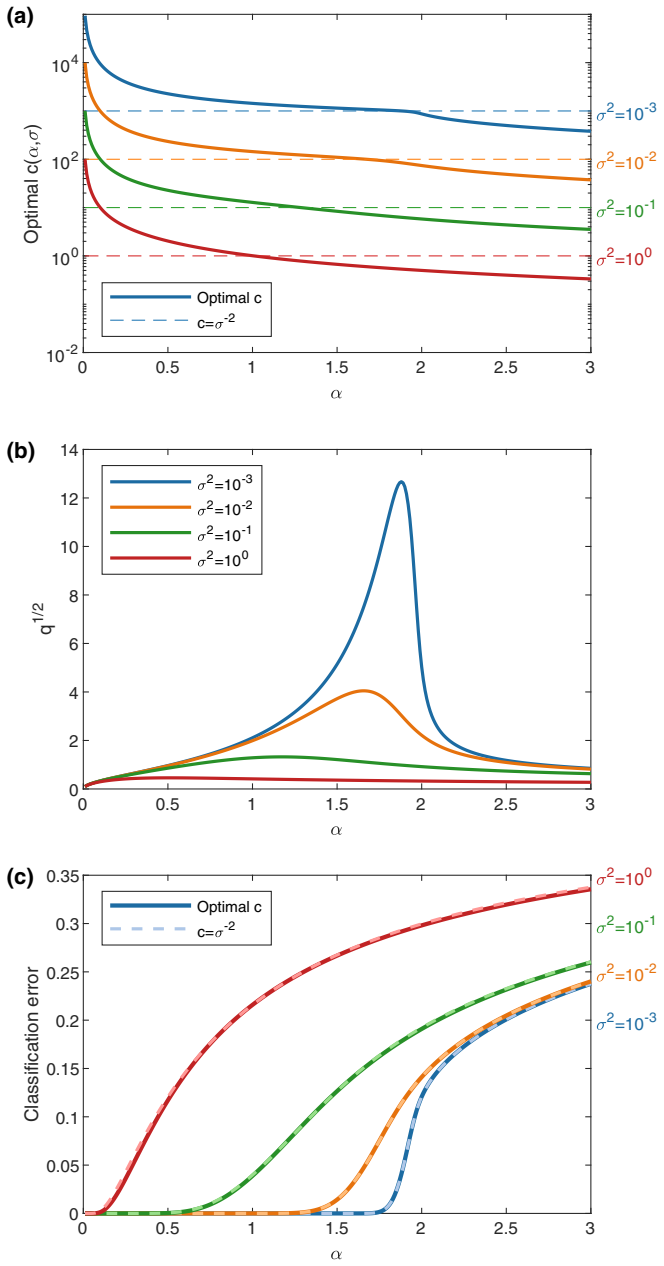

FIG. 3. The optimal choice of c in soft classification of points. (a) The optimal choice of c (y axis, log scale) for different values of α (x axis) and levels of noise σ^2 (color coded). Compares the optimal choice c^* (solid lines) and the canonical choice $c = \sigma^{-2}$ (dashed lines). (b) The weights' norm $q^{1/2}$ (y axis) for different values of α (x axis) and levels of noise σ^2 (color coded) when using the optimal value of c. (c) Classification error (y axis) for different values of α (x axis) and levels of noise σ^2 (color coded). Compares the optimal choice c^* (solid lines) and the canonical choice $c = \sigma^{-2}$ (dashed lines).

become

$$h^\mu(\vec{S}) = y^\mu \boldsymbol{w} \cdot \boldsymbol{x}^\mu(\vec{S}) = v_0^\mu + \vec{S} \cdot \vec{v}^\mu. \tag{21}$$

The classic soft classification formalism [3], called here point-slack SVM, uses one slack variable per sample. It is usually inapplicable for manifold classification as the number of samples may be infinite. Thus we consider two simple alternatives which allow for soft classification of manifolds; both require only a single slack variable per manifold. In several specific cases where the point-slack formalism can be used, it is compared with those formalisms.

Center-slack method A naive approach for the classification of manifolds is to assume the soft classifier is learned using only the manifolds' centers and then evaluated on the entire manifolds. Formally, soft classification using center slacks is defined by weights $\boldsymbol{w} \in \mathbb{R}^N$ and slack variables $\vec{s} \in \mathbb{R}^P$ such that the central fields obey for all $\mu \in [1, \ldots, P]$

$$v_0^\mu = y^\mu \boldsymbol{w} \cdot \boldsymbol{u}_0^\mu \geqslant 1 - s^\mu. \tag{22}$$

Given a regularization parameter $c \geqslant 0$ the optimal classifier is defined by the Lagrangian

$$L = \|\boldsymbol{w}\|^2/N + c\|\vec{s}\|^2/N \text{ s.t. } \forall \mu \; v_0^\mu \geqslant 1 - s^{\mu\cdot}. \tag{23}$$

Using this method the manifold structure is not used during training, so the weights' norm and field distribution (with respect to the centers) are given by points classification theory from previous section. However, an evaluation of classification errors on the manifold would require additional assumptions on the manifold.

Manifold-slack method The previous method uses a slack variable to constrain the mean of the fields on the manifold. A natural alternative would be to constrain the minimal field on the manifold. Using the field definition $h^\mu(\vec{S})$, soft classification using manifold slacks is defined by weights $\boldsymbol{w} \in \mathbb{R}^N$ and slack variables $\vec{s} \in \mathbb{R}^P$ where the minimal fields obey for all $\mu \in [1, \ldots, P]$

$$h_{\min}^\mu \doteq \min_{\vec{S} \in \mathcal{M}^\mu} h^\mu(\vec{S}) \geqslant 1 - s^\mu. \tag{24}$$

That is, given a regularization parameter $c \geqslant 0$ the optimal classifier is defined by the Lagrangian

$$L = \|\boldsymbol{w}\|^2/N + c\|\vec{s}\|^2/N \text{ s.t. } \forall \mu \; h_{\min}^\mu \geqslant 1 - s^\mu. \tag{25}$$

Figure 4 illustrates soft classification of points (or manifold centers, as noted above), spheres, and general manifolds. In what follows we first discuss spheres, then extend the discussion to general manifolds.

C. Soft classification of spheres

A D-dimensional sphere of radius R in $\mathbb{R}^N$ is defined:

$$\boldsymbol{x}^\mu(\vec{S}) = \boldsymbol{u}_0^\mu + R \sum_{l=1}^{D} S_l \boldsymbol{u}_l^\mu \text{ s.t. } \|\vec{S}\| \leqslant 1. \tag{26}$$

As in the case of points we would analyze the classification problem for random labels $\vec{y} \in \{\pm 1\}^P$ and random axes $\boldsymbol{u}_{li}^\mu \sim \mathcal{N}(0, 1/N)$, i.e., again scaling $\|\boldsymbol{u}_l^\mu\| \approx 1$.

1. Center slack

Using center slacks the classifier properties are given by the theory of soft classification of points, self-consistent Eqs. (7) and (8), and the distribution of the fields on the centers follows Eq. (13).

The classification error on the sphere is defined $\varepsilon = P(v_0 + R\sum_l v_l S_l \leqslant 0)$ but as $v_l = y\boldsymbol{w} \cdot \boldsymbol{u}_l$ where $\boldsymbol{w}$ is independent of

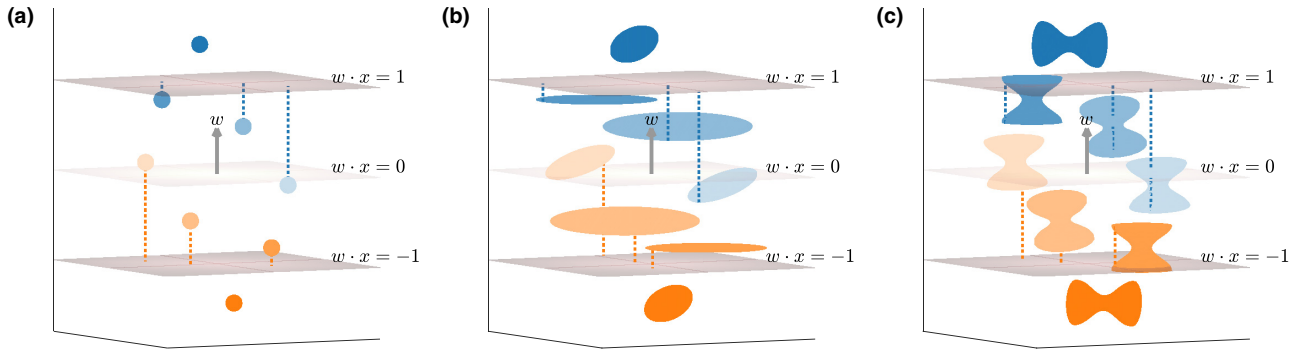

FIG. 4. Illustration of soft classification of points, spheres, and general manifolds. A weight vector $\boldsymbol{w}$ (gray arrow) defines the signed fields $\boldsymbol{w} \cdot \boldsymbol{x}$ on the manifolds being classified and satisfies $y\boldsymbol{w} \cdot \boldsymbol{x} \geqslant 1 - s$. The light gray hyperplane depicts the decision boundary $\boldsymbol{w} \cdot \boldsymbol{x} = 0$; points above it are labeled $+1$ and below it -1. The dark gray hyperplanes depict the boundaries $\boldsymbol{w} \cdot \boldsymbol{x} = \pm 1$. The length of each manifold's slack is indicated by a dashed line from the manifold point with the minimal field $y\boldsymbol{w} \cdot \boldsymbol{x}$ to the hyperplane $\boldsymbol{w} \cdot \boldsymbol{x} = y$. Each panel depicts the classification of four blue manifolds (target label is $+1$) against four orange manifolds (target label is -1). The blue and orange manifolds are symmetrically positioned for illustration purposes only. Manifolds are numbered from darkest to lightest. (a) Classification of points: the first point is in the interior $y\boldsymbol{w} \cdot \boldsymbol{x} > 1$, has $s = 0$; the second and third points have nonzero slack $0 < s < 1$, are classified correctly; the fourth point is below the decision boundary $y\boldsymbol{w} \cdot \boldsymbol{x} < 0$, corresponds to an error, and has $s > 1$. (b) Classification of spheres: the first sphere is in the interior $y\boldsymbol{w} \cdot \boldsymbol{x} > 1$, the second sphere is fully embedded within the hyperplane $y\boldsymbol{w} \cdot \boldsymbol{x} = 1 - s$, and the third and fourth spheres touching the hyperplane $y\boldsymbol{w} \cdot \boldsymbol{x} \geqslant 1 - s$ with the minimal field above 0 for the third and below 0 for the fourth. (c) Classification of general manifolds: the first manifold is in the interior $y\boldsymbol{w} \cdot \boldsymbol{x} > 1$, the second manifold has a face embedded within the hyperplane $y\boldsymbol{w} \cdot \boldsymbol{x} = 1 - s$, and the third and fourth manifolds touching the hyperplane $y\boldsymbol{w} \cdot \boldsymbol{x} \geqslant 1 - s$ with the minimal field above 0 for the third and below 0 for the fourth.

$\boldsymbol{u}_l$ in this case, we have that $v_l \sim \mathcal{N}(0, q)$, and as $\|\vec{S}\| = 1$ on the sphere $R\sum_l v_l S_l \sim \mathcal{N}(0, qR^2)$. If we assume Gaussian noise $\mathcal{N}(0, \sigma^2/N)$ is added independently for each sample component, as we have done for points, we have noise $\mathcal{N}(0, q(\sigma^2 + R^2))$ at the fields. Thus the error is given by $\varepsilon = P(v_0 + \sqrt{(\sigma^2 + R^2)q}\eta \leqslant 0)$ where η is a standard Gaussian, or equivalently

$$\varepsilon = \langle H(v_0/\sqrt{(\sigma^2 + R^2)q})\rangle_{v_0}, \tag{27}$$

where surprisingly, the dimensionality D of the spheres plays no role in this setting.

We conclude that soft classification of spheres of radius R using center slacks with noise level of σ^2 is equivalent to soft classification of points with effective noise $\sigma^2_{eff} = \sigma^2 + R^2$. Several corollaries can be made from the analysis of points, by using the effective noise σ^2_{eff} instead of σ^2. First, when $(\sigma^2 + R^2)q \ll 1$ we expect a good SNR approximation $\varepsilon \approx H(\mathcal{S})$ using

$$\mathcal{S} = ck/\sqrt{q[1 + (\sigma^2 + R^2)(1 + ck)^2]}. \tag{28}$$

Figure 5(a) shows the resulting error when sampling from the sphere (i.e., $\sigma = 0$) for different values of R, and Fig. S6 compares the theory to the error measured empirically. Second, the optimal choice of c is then given by Eq. (18), as well as the "canonical choice"

$$c \approx 1/(\sigma^2 + R^2). \tag{29}$$

Contrary to the result from classification of points, due to the contribution of R, here the optimal choice for c is finite even for $\sigma = 0$, as illustrated in Fig. 5(b).

2. Manifold slack

We now consider soft classification of the entire manifold, that is, $h^\mu_{\min} \geqslant 1 - s^\mu$, thus generalizing the analysis of max-margin classifiers for spheres [6]. For spheres the point with the "worst" field, or minimal overlap with $\boldsymbol{w}$, is given by $\vec{S} = -\hat{v}$ (where $\hat{v} = \vec{v}/\|\vec{v}\|$), and hence a necessary and sufficient condition for the soft classification of the entire sphere is given by $v^\mu_0 - R\|\vec{v}^\mu\| \geqslant 1 - s^\mu$.

Replica theory This observation allows us to write an expression for the volume $V(L, c)$ of solutions achieving a target value of the loss L:

$$V(L, c) = \int d^N\boldsymbol{w} \int d^P\vec{s}\,\delta(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL) \tag{30}$$

$$\cdots \prod_\mu^P \delta(v^\mu_0 - R\|\vec{v}^\mu\| - h^\mu)\Theta(h^\mu - 1 + s^\mu). \tag{31}$$

A replica analysis yields the following relation between L, α and the same order parameters q, k (i.e., defined exactly as in the case of points) when the volume of solutions vanishes (see Appendix A 7):

$$L/q = \frac{k-1}{k} + \frac{\alpha}{k}\int D^D\vec{t} \int Dt_0 F(\vec{t}, t_0), \tag{32}$$

$$F(\vec{t}, t_0) = \min_{v_0 - R\|\vec{v}\| \geqslant 1/\sqrt{q}} \left\{ \|\vec{v} - \vec{t}\|^2 + \frac{ck}{1 + ck}(v_0 - t_0)^2 \right\}, \tag{33}$$

where $q = \|\boldsymbol{w}\|^2/N$ and $Dt_0 = dt_0 e^{-t_0^2/2}/\sqrt{2\pi}$ so $\vec{t}, t_0$ are $D + 1$ Gaussian variables representing the quenched noise in the solution, due to the variability of the labels $\{y^\mu\}$

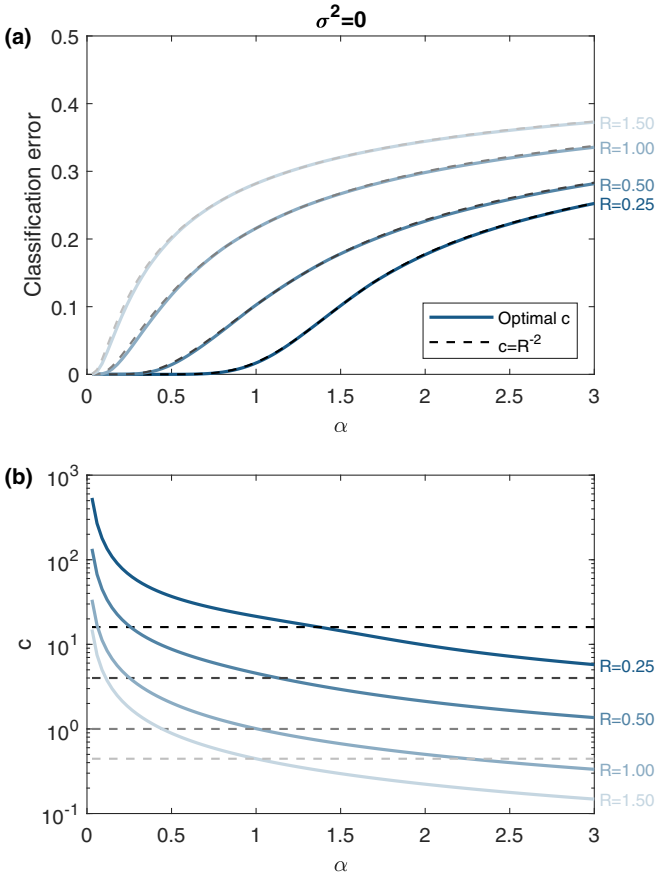

FIG. 5. Soft classification of spheres using center slacks. (a) Classification error (y axis) for different values of α (x axis) and R (color coded), without noise $\sigma^2 = 0$, using the optimal choice of c (solid lines) and the canonical choice $c = R^{-2}$ (dashed lines). (b) The optimal choice of c (y axis, log scale) for different values of α (x axis) and R (color coded), without noise $\sigma^2 = 0$. The canonical choice $c = R^{-2}$ is indicated by the dashed horizontal lines. Those results are independent of D; see main text.

and the manifolds' axes $\{\boldsymbol{u}_l^\mu\}$. Note that for $D = 0$, $F(t_0) = \frac{ck}{1+ck}\alpha_0^{-1}(1/\sqrt{q})$ so we recover Eq. (6).

Solving the inner problem [Eq. (33)] using Karush-Kuhn-Tucker conditions [19] (KKT) allows us to describe the joint distribution of v_0, $v = \|\vec{v}\|$, and s conditioned on $t_0, t = \|\vec{t}\|$ at different solution regimes (see Appendix A 8):

(1) "Interior" regime: the entire sphere is classified correctly with $h > 1$ and a margin larger than $1/\sqrt{q}$ from the hyperplane $h = 0$; in this regime the slack is not utilized $s = 0$ and the solution satisfies $v_0 = t_0$, $v_l = t_l$ so that $F = 0$. This regime is in effect for $1/\sqrt{q} + Rt \leqslant t_0 \leqslant \infty$.

(2) "Touching" regime: the tip of the sphere touches the hyperplane $h = 1 - s$; in this regime v_0, v, s have nontrivial values. This regime is in effect for $1/\sqrt{q} - \frac{1+ck}{ck}t/R \leqslant t_0 \leqslant 1/\sqrt{q} + Rt$.

(3) "Embedded" regime: the entire sphere is within the hyperplane $h = 1 - s$; in this regime $v = 0$ but v_0, s have nontrivial values. This regime is in effect for $-\infty < t_0 \leqslant 1/\sqrt{q} - \frac{1+ck}{ck}t/R$.

The same KKT analysis also provides the minimization value $F(t_0, t)$ achieved at each regime, so that denoting $f(R, D, ck, q) = \int D^D\vec{t} \int Dt_0 F(\vec{t}, t_0)$ and the chi distribution with D degrees of freedom $\chi_D(t) = \frac{2^{1-D/2}}{\Gamma(D/2)} t^{D-1} e^{-t^2/2}\, dt$:

$$f(R, D, ck, q) = \int \chi_D(t) \int_{-\infty}^{1/\sqrt{q} - \frac{1+ck}{ck}t/R} Dt_0 \left[\frac{ck}{1+ck}(1/\sqrt{q} - t_0)^2 + t^2\right] \tag{34}$$

$$+ \int \chi_D(t) \int_{1/\sqrt{q} - \frac{1+ck}{ck}t/R}^{1/\sqrt{q}+Rt} Dt_0 \frac{ck}{1+ck(1+R^2)}(1/\sqrt{q} + Rt - t_0)^2, \tag{35}$$

and the mean-field equation becomes

$$L/q = \frac{k-1}{k} + \frac{1}{k}\alpha f(R, D, ck, q). \tag{36}$$

Self-consistent equations Assuming the optimal loss L^* satisfies saddle-point conditions $0 = \frac{\partial L}{\partial q} = \frac{\partial L}{\partial k}$, we have two self-consistent equations for k, q, similar to those found in the case of points:

$$1 = \alpha f - \alpha k \frac{\partial}{\partial k} f, \tag{37}$$

$$1 - k = \alpha f + \alpha q \frac{\partial}{\partial q} f. \tag{38}$$

See the concrete form in A 9. Those equations can be solved numerically to predict the weights' norm (Algorithm 2). This prediction is compared to the norm observed in simulations (i.e., by finding the optimal weights for classification of spheres, Algorithm 3). Figure 6 shows the resulting q, k for specific values of R, D (and additional ones are presented in Fig. S8); $q(\alpha)$ has a single peak, increasing from 0 to a finite value at the peak, then decreasing monotonically, while $k(\alpha)$ decrease monotonically from 1 to 0. We note that while for spheres ck no longer corresponds exactly to the ratio between the two parts of the optimization target, its interpretation as a measure of the contribution of the weights is maintained.

The mean-field equations can be simplified when considering several interesting limits (see Appendix A 10). As in the case of points, in the limit $c \to \infty$, we find a different behavior below and above α_C^{Hard}, the max-margin capacity. For $\alpha < \alpha_C^{\text{Hard}}$ we have that k is finite while ck diverges, with Eq. (37) becoming the mean-field equation from max-margin classification [6], and the underlying Lagrangian is given by

$$L = \|\boldsymbol{w}\|^2/N \;\; \text{s.t.}\; \forall\mu \; h_{\min}^\mu \geqslant 1. \tag{39}$$

On the other hand, for $\alpha > \alpha_C^{\text{Hard}}$ we have that k approaches 0 while q and $K = \lim_{c\to\infty} ck$ are finite, with the underlying Lagrangian

$$L = \|\vec{s}\|^2/N \;\; \text{s.t.}\; \forall\mu \; h_{\min}^\mu \geqslant 1 - s^\mu. \tag{40}$$

A second interesting limit is $\alpha \to 0$. In this limit we expect the order parameters to behave as in the case of points, $q \to 0$ and $k \to 1$. We find that for small α the self-consistent equations are simplified, and for $\alpha \ll 1$ we have the approximations $k \approx 1 - \alpha(1 + D)$ and $q \approx \alpha(ck)^2/(1 + ck)^2$ (see Fig. 6 and Fig. S8 where those approximations are marked).

Phase transition An analysis of the mean-field equations reveals that for spheres (unlike points) there is a finite value of α where $q \to 0$, and above which the self-consistent

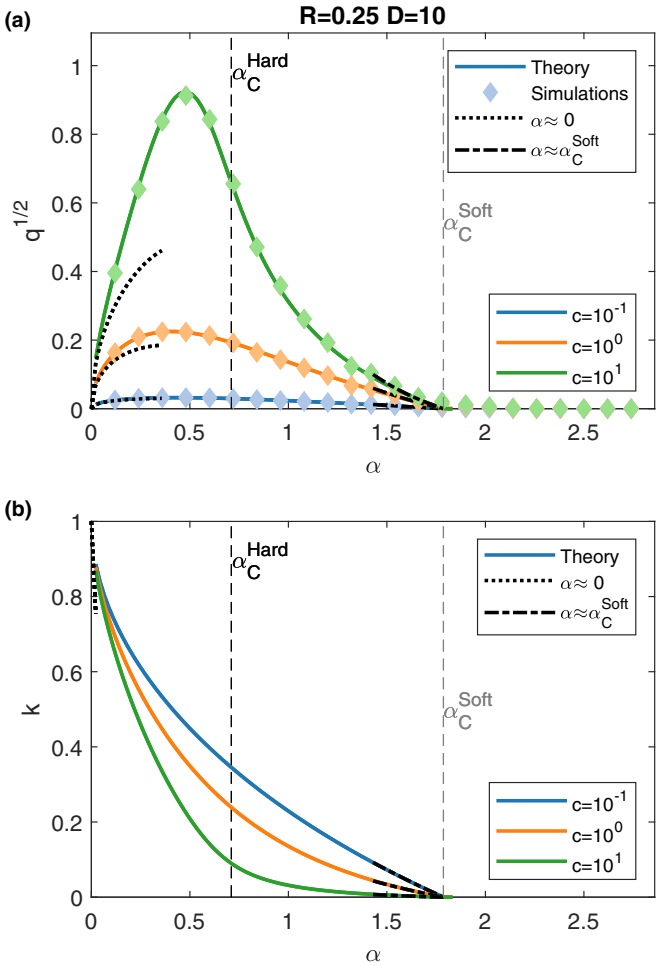

FIG. 6. Order parameters in soft classification of spheres using manifold slacks. (a) The weights' norm $q^{1/2}$ (y axis) for different values of α (x axis), and choices of c (color coded), for radius $R = 0.25$ and dimension $D = 10$. Compares theory results (solid lines) to simulation results (diamonds). (b) The order parameter k (y axis) for different values of α (x axis) and choices of c (color coded). (a, b) Theory for the limits of $\alpha \to 0, \alpha \to \alpha_C^{\text{Soft}}$ is marked as black dotted and dash-dot lines, respectively.

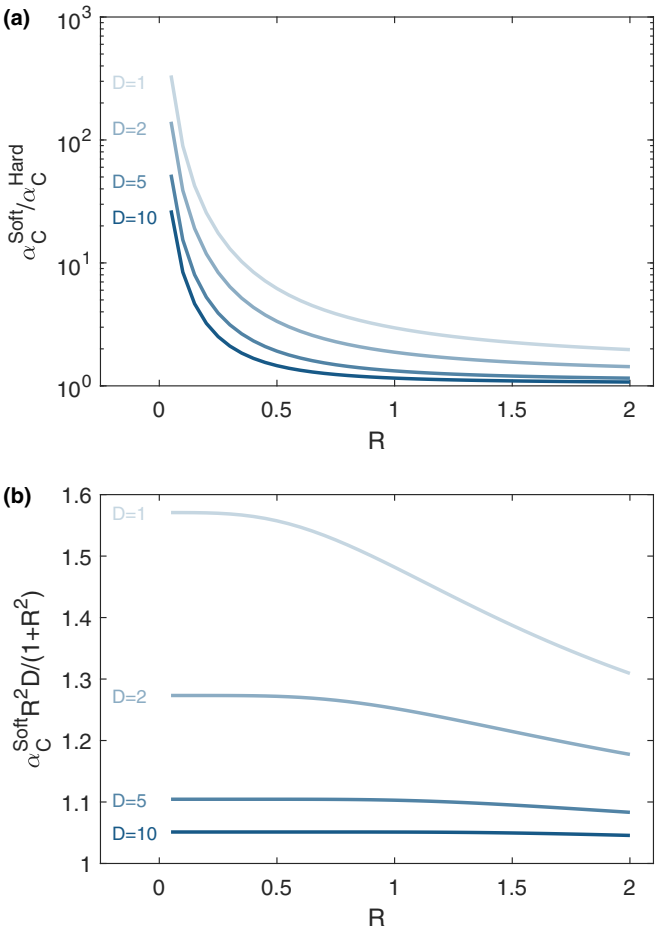

FIG. 7. Capacity in manifold-slack classification of spheres. (a) The ratio between α_C^{Soft} and α_C^{Hard} (y axis, log scale) for different values of R (x axis) and D (color coded). (b) The ratio between α_C^{Soft} and Eq. (43) approximation (y axis) for different values of R (x axis) and D (color coded).

equations cannot be solved for k, q (see Figs. 6 and S8). The corresponding simulation results indicate that when the theory equations cannot be solved the optimal classifier is $\boldsymbol{w} = \mathbf{0}$, that is $q = 0$, with all the slack variables saturating at $\vec{s} \equiv 1$. Thus, soft margin classification problems always have a solution, unlike max-margin problems, but above a certain value of α this is the trivial solution. The critical value for α can be found by assuming that both $k, \sqrt{q} \ll 1$; using a scaling of $x = ck/\sqrt{q}$ we get that $\alpha = \alpha_C$ would satisfy (see Appendix A 11)

$$\alpha_C^{-1} = \int_0^{xR} \chi_D(t)t^2 + xR \int_{xR}^{\infty} \chi_D(t)t, \tag{41}$$

$$x = \left(1 + R^2 \int_{xR}^{\infty} \chi_D(t)\right)^{-1} R \int_{xR}^{\infty} \chi_D(t)t, \tag{42}$$

where x is the self-consistent solution of Eq. (42).

Surprisingly, the critical value is independent of c and we denote it α_C^{Soft}, as a soft analog of the max-margin capacity α_C^{Hard} [6]. Notably, the former is always larger $\alpha_C^{\text{Soft}} \geqslant \alpha_C^{\text{Hard}}$, as shown in Fig. 7(a).

For $R \to 0$ we have that $x = R\int_0^{\infty} \chi_D(t)t = R\sqrt{2}\Gamma(\frac{D}{2} + \frac{1}{2})/\Gamma(\frac{D}{2})$ and $\alpha_C^{-1} = x^2$. Thus, for small R, the critical value α_C^{Soft} diverges as R^{-2} (and in the limit of points there is no phase transition). Conversely, for $R \to \infty$ we have $x \approx 0$ and $\alpha_C^{\text{Soft}} = D^{-1}$, whereas in this limit $\alpha_C^{\text{Hard}} = (D + 1/2)^{-1}$ [6]. Intuitively, in both cases $\boldsymbol{w}$ must be perpendicular to the PD manifold axes; for soft classification this implies just $N > PD$ or $\alpha < D^{-1}$, while for max-margin classification due to the finite capacity when classifying the centers this means $P/(N - PD) < 2$ or $\alpha < (D + 1/2)^{-1}$.

The existence of a sharp transition in the manifold-slack problem is the result of the thermodynamic limit. For small N, the existence of a solution at any given α depends on the particular labels realization. As N increases, the probability of having a solution approaches 1 for $\alpha < \alpha_C$, and 0 for $\alpha > \alpha_C$ (Fig. S7).

Phase transition for large D regime When $D \gg 1$ the phase transition Eqs. (41) and (42) implies a simple expression for

capacity:

$$\alpha_C^{\text{Soft}} \approx (1+R^2)/R^2 D. \tag{43}$$

Figure 7(b) compares this approximation to the full expression for different values of R, D; as observed, this approximation is reasonable for large D independently of the value of R. In this regime the max-margin capacity is given by [6]

$$\alpha_C^{\text{Hard}} \approx (1+R^2)\alpha_0(R\sqrt{D}). \tag{44}$$

Phase-transition intuition To gain some intuition for why a phase transition is to be expected for manifold slack, we need to consider the distribution of slack values. As for points, the mean-field theory provides the full distribution of the fields and slack variables (see Appendix A 12). Figure S9 compares the theoretical slack distribution to the histogram of the values observed in simulations. We note that the slack distribution depends on q both for the mean and the variance; decreasing q pushes the slack distribution toward a δ-function at 1 when $q = 0$. Now consider how manifold slack is compared to center slack. From the theory of classification of points, the loss in classification using center slacks monotonically increases in α and tends asymptotically (from below) toward $L = c\alpha$. As this is the loss achieved by the trivial solution, reaching it at a finite α corresponds to the phase transition. For large values of α, weights trained on manifold centers have small q and slack values near 1; those achieve $h_{\min} \approx v_0 - \sqrt{q}R\sqrt{D}$. Thus if those weights were used by the manifold slack, the slacks would need to increase by $\sqrt{q}R\sqrt{D}$, pushing their mean above 1. To avoid this, the loss is reduced by decreasing q, pushing it toward 0. Below we use this intuition to speculate on whether other soft classification formalisms would introduce a phase transition (see discussion).

Classification errors Next we use the fields and slack distributions to calculate the probability of classification errors. In the framework of manifold slacks it is natural to consider the probability of error anywhere on the manifold, or equivalently the fraction of manifolds where the worst point is misclassified. This is the fraction of slack variables that are larger than 1, i.e., $\varepsilon_{tr}^{\text{manifold}} = P(s \geqslant 1)$, which can be evaluated from the slack distribution. This entire-manifold classification error is given by $\varepsilon_{tr}^{\text{manifold}} = H(\mathcal{S}^{\text{manifold}})$ for $\mathcal{S}^{\text{manifold}}$ defined in A 13.

A different kind of error is the probability of classification error on uniformly sampled points from the sphere, that is, $\varepsilon_{tr}^{\text{sample}} = P(h < 0)$, similar to the error considered above for center slacks. These fields can be written as $h = v_0 + Rvz$, where $z = \cos(\theta)$ for θ the angle between the weight vector and the point on the sphere, v_0 and $v = \|\vec{v}\|$ are the projections of the weight vector on the center and the sphere subspace. Thus, $\varepsilon_{tr}^{\text{sample}} = P(v_0 + Rvz < 0)$, where the joint distribution of v_0, v is given by theory, and for a uniform sampling from a sphere $z \in [-1, 1]$ has a bell-shaped distribution (see Appendix A 13):

$$P(z) = \frac{1}{\sqrt{\pi}}(1-z^2)^{\frac{D-3}{2}}\Gamma\left(\frac{D}{2}\right)\Big/\Gamma\left(\frac{D-1}{2}\right) \tag{45}$$

with moments $\langle z \rangle = 0$ and $\langle \delta z^2 \rangle = 1/D$. In this setting classification error monotonically decreases with c so the optimal value of $\varepsilon_{tr}^{\text{sample}}$ is achieved for $c = \infty$.

We now consider the classification error of points on the sphere in the presence of noise, where the classifier is trained on the entire manifold (i.e., with no noise), and tested on noisy samples from the manifold. Assuming Gaussian noise $\mathcal{N}(0, \sigma^2/N)$ is added to each component of manifold samples, the fields are affected by noise $\mathcal{N}(0, \sigma^2 q)$. Thus the probability of error in a sample is given by $P(h + \sigma\sqrt{q}\eta < 0)$ where η is standard Gaussian, and equivalently

$$\varepsilon_g^{\text{sample}} = \left\langle H\left(\frac{v_0 + Rzv}{\sigma\sqrt{q}}\right)\right\rangle_{v_0, v, z}. \tag{46}$$

Large D regime The regime of spheres with $D \gg 1$ is important as real-world manifolds are expected to be high-dimensional, and in this regime it is possible to derive an SNR approximation of Eq. (46).

When $R \sim O(1)$, α_C^{Soft} is close to α_C^{Hard} (see Fig. 7). Thus in this regime the benefit of soft classification, in terms of the range of valid solutions, is small. On the other hand, when $R\sqrt{D} \sim O(1)$, α_C^{Soft} can be much larger than α_C^{Hard} [Fig. 7(a)], and thus we focus on this regime in our analysis of classification errors.

To derive an SNR approximation we assume that in this regime $v_0 + Rzv$ is approximately Gaussian, and that only the "touching" regime contributes to the error, thus substituting the values of v_0, v derived from the mean-field theory in that regime. The resulting SNR is provided in Appendix A 13.

Importantly, from this analysis we can calculate the limiting behavior of the SNR. In the $\alpha \to 0$ limit the error anywhere on the manifold scales as $\lim_{\alpha\to 0} \varepsilon_{tr}^{\text{manifold}} = H(ck/\sqrt{q})$, and using the order parameters in this limit leads to

$$\lim_{\alpha\to 0} \varepsilon_{tr}^{\text{manifold}} = H[(1+c)/\sqrt{\alpha}], \tag{47}$$

which is exactly the scaling for classification of the center points alone [$\varepsilon_{tr}^{\text{centers}}$, Eq. (17) with $\sigma^2 = 0$]. Thus in this regime (i.e., $N \to \infty$) the manifold structure does not affect the classification error, and furthermore the error in classification of the entire sphere is the same as the error in classification of samples $\varepsilon_{tr}^{\text{sample}}$, as the former is bounded between the two classification errors $\varepsilon_{tr}^{\text{centers}} \leqslant \varepsilon_{tr}^{\text{sample}} \leqslant \varepsilon_{tr}^{\text{manifold}}$.

On the other hand, in the $\alpha \to \alpha_C^{\text{Soft}}$ limit, from the scaling of k, q in this limit the error in classifying the entire manifold saturates, but not the error classifying samples (see Appendix A 13):

$$\lim_{\alpha\to\alpha_C^{\text{Soft}}} \varepsilon_{tr}^{\text{manifold}} = H(0) = 1/2, \tag{48}$$

$$\lim_{\alpha\to\alpha_C^{\text{Soft}}} \varepsilon_g^{\text{sample}} = H\left(\frac{R\sqrt{D}}{1+R^2}\frac{1}{\sqrt{1+\sigma^2}}\right). \tag{49}$$

Thus the theory predicts that errors at the phase transition are independent of c and jump from this finite value to 0.5 (in simulations using a finite N this transition is smoothed, as already discussed above).

Figure 8(a) presents both types of training errors and their dependence on α and c at specific values of R, D, demonstrating that they are monotonically decreasing with c and monotonically increasing with α. Unlike the training error, in the presence of noise the test error is not monotonic in both α [Fig. 8(b)] and c [Fig. 8(c)]. Thus error is minimized for a

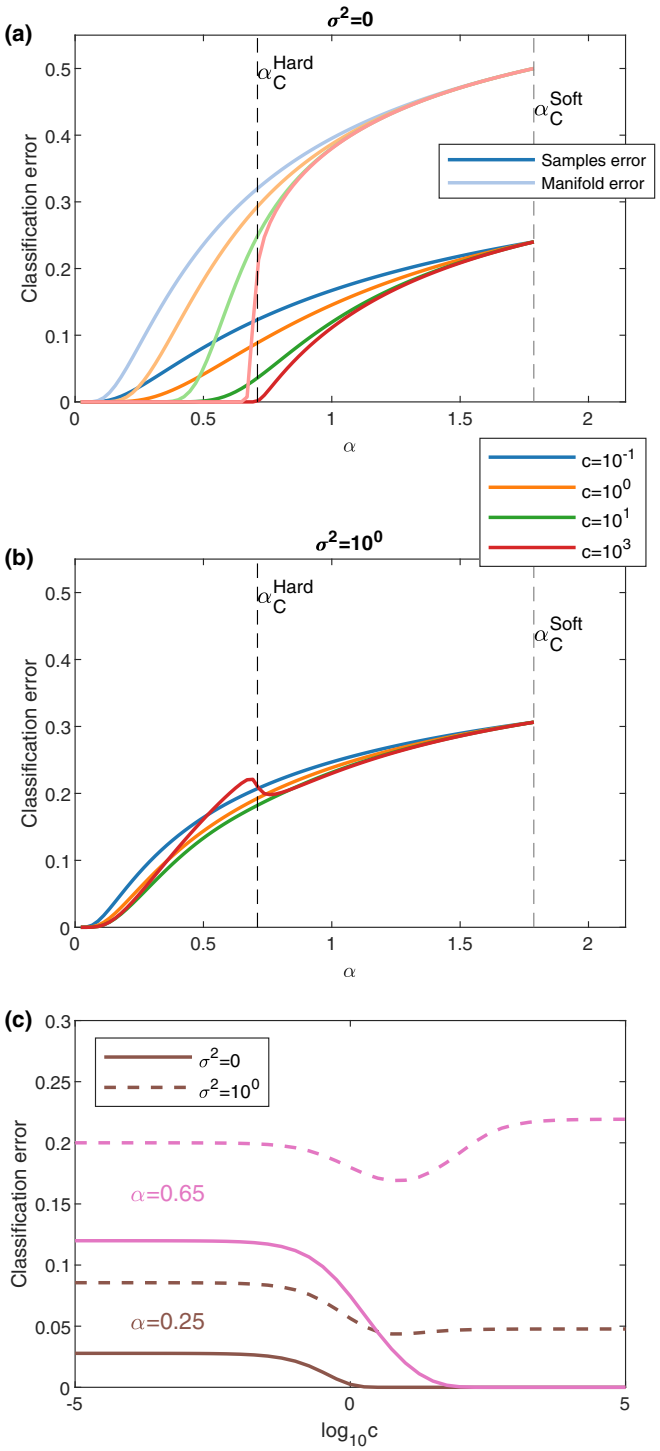

FIG. 8. Errors in soft classification of spheres using manifold slacks. Results for spheres of radius $R = 0.25$ and dimension $D = 10$. (a) Classification error without noise (y axis) for different values of α (x axis) and choices of c (color coded). Compares samples' classification error (dark lines) and entire-manifold classification error (light lines). (b) Classification error at noise level $\sigma^2 = 1$ (y axis) for different values of α (x axis) and choices of c (color coded). (c) Classification error (y axis) for different choices of c (x axis, log scale), for several values of α (color coded), levels of noise σ^2 (solid/dashed lines).

finite value of c, which depends on both the noise level σ and the load α.

Agreement of the theory with empirical simulations is presented for different parameter values and choices of c in Fig. S10 for the training error, and similarly in Fig. S11 for the test error. Thus theory can be used to choose the optimal value of c. Figure S12 presents the optimal value of c for different values of α and levels of noise, demonstrating a nontrivial behavior for manifold slacks, unlike the monotonic behavior predicted by theory for center slacks.

Comparison with other methods Comparing the performance of the manifold-slack method with other methods requires optimization of the regularization value c independently for each method. When there is no noise, below max-margin capacity $\alpha < \alpha_C^{\text{Hard}}$, the optimal choice of c is infinite such that manifold-slack classification converges to max-margin classification. However, in the presence of noise the optimal value of c is finite and using manifold slacks reduces classification error relative to max-margin classification [Fig. 9(a)]. While the manifold-slack method is strictly better than the max-margin method due to choosing from a larger pool of classifiers, the improvement is usually small and is achieved toward α_C^{Hard} (see Fig. S13).

A systematic comparison of the manifold-slack and center-slack methods finds that manifold slacks are better for small α values, with notable benefits at larger R and smaller σ values [see Figs. 9(b) and 9(c)]. Intuitively, when the noise is small, manifold slacks may achieve near-zero error at a range of α values, while center-slack performance depends on R as a noise term and thus may be order 1 when R is order 1. For larger α values the performance of center slacks surpasses that of manifold slacks, and finally above α_C^{Soft} only the center-slack method is a viable option.

As noted above, the point-slack method cannot in general be used for classification of manifolds with an infinite number of points. However, for classification of line segments (i.e., spheres with $D = 1$), a correct classification of the $2P$ end points is enough to classify the entire line. Figure S14 compares manifold slack with point-slack classification of the $2P$ end-points, both using the optimal choice of c for a given level of noise. The performance of point-slack SVM is usually close to that of the manifold-slack method, but provides a significant improvement toward α_C^{Soft}. The line segments case demonstrates a striking contrast between those alternatives. Using the manifold-slack method (with P slack variables) there is a phase transition where the nontrivial classifier vanishes at a finite α, as expected from spheres classification theory. But there is no such transition using the point-slack method (with $2P$ slack variables), as expected from the point-slack theory [compare the weights' norms in Figs. S14(a) and S14(b)].

D. Soft classification of general manifolds

We now consider the more general case, which is relevant for applications, where the above analysis of both points and spheres would serve as a stepping stone toward theoretic understanding of general manifolds.

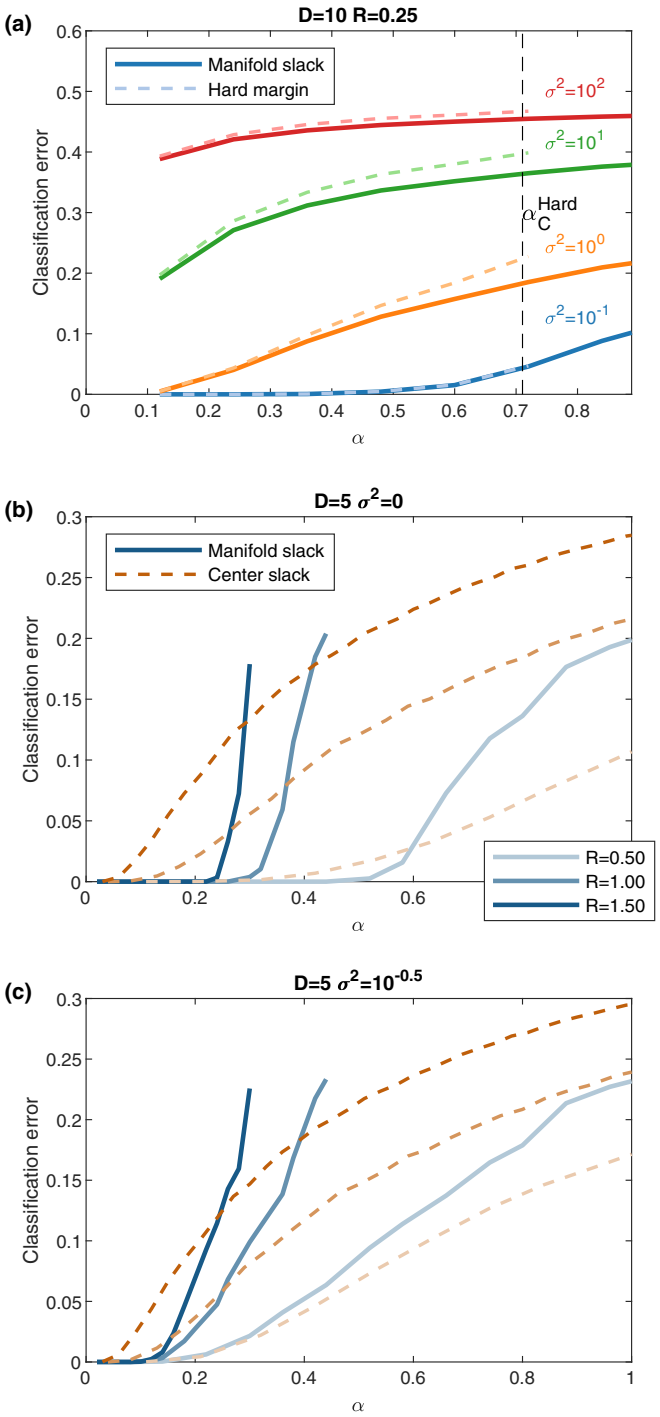

FIG. 9. Comparison of classification errors for spheres using different methods. (a) Classification error (y axis) using manifold slacks (at the optimal choice of c, solid lines) or max-margin classification (dashed lines) at different values of α (x axis) for radius $R = 0.25$ and dimension $D = 10$. Compares simulation results at different noise levels (color coded). (b, c) Classification error using the optimal choice of c (y axis) for different values of α (x axis) and values of R (color coded), for dimension $D = 5$. Compares simulation results of manifold-slack classifiers (solid lines) and center-slack classifiers (dashed lines), without noise (b) and with noise (c).

1. Center slack

The center-slack method is straightforward to generalize to general manifolds, with the centers defined per our definition of a general manifold [$\boldsymbol{u}_0$ in Eq. (20)]. A classifier trained on the centers would have a norm per points classification theory [Eqs. (7) and (8)], and central field distribution per Eq. (13).

The probability of classification error for a point on the manifold $\boldsymbol{x}(\vec{S})$ would be $\varepsilon(\vec{S}) = P(v_0 + \vec{S}\cdot\vec{v} \leqslant 0)$ with $\vec{S}\cdot\vec{v} \sim \mathcal{N}(0, q\|\boldsymbol{x}(\vec{S}) - \boldsymbol{u}_0\|^2)$. A calculation of classification error on a general manifold requires to make further assumptions on the sampling of $\vec{S} \in \mathcal{M}$ (see discussion). However, for the simple case of uniform sampling from a point-cloud manifolds where $\boldsymbol{x}_m = \boldsymbol{u}_0 + \delta\boldsymbol{x}_m$ for $min[1\ldots M]$ we have that

$$\varepsilon = \frac{1}{M}\sum_{m=1}^{M}\langle H(v_0/\sqrt{(\sigma^2 + \|\delta\boldsymbol{x}_m\|^2)q})\rangle_{v_0}, \tag{50}$$

where σ^2/N is the variance of Gaussian noise added to each component, which generalize Eq. (27) from spheres, with the empirical $\|\delta\boldsymbol{x}_m\|^2$ taking the role of R^2. Furthermore, when the number of samples is large we expect self-averaging:

$$\varepsilon = \langle H(v_0/\sqrt{(\sigma^2 + \hat{R}^2)q})\rangle_{v_0} \tag{51}$$

for $\hat{R}^2 = \frac{1}{M}\sum_{m=1}^{M}\|\delta\boldsymbol{x}_m\|^2$ the total variance of the manifold points. Figure S15 compares the full theory [Eq. (50)] and the approximation [Eq. (51)] to empirical measurement of the error using center slacks.

2. Manifold slack

Replica theory Generalizing the mean-field theory of spheres to general manifolds, the theory implies that Eq. (32) is unmodified while Eq. (33) becomes

$$F(\vec{t}, t_0) = \min_{v_0 + g(\vec{v}) \geqslant 1/\sqrt{q}}\left\{\|\vec{v} - \vec{t}\|^2 + \frac{ck}{1+ck}(v_0 - t_0)^2\right\}, \tag{52}$$

where $g(\vec{v}) = \min_{S\in\mathcal{M}}\vec{v}\cdot\vec{S}$ is a called the "support function." To characterize the solution of $F(\vec{t}, t_0)$ using KKT conditions, we formally define "anchor points" as the subgradient of the function (as in [7])

$$\tilde{S}(\vec{v}) = \frac{\partial}{\partial v}g(\vec{v}), \tag{53}$$

and when the support function is differentiable, the subgradient is unique and is equivalent to the gradient:

$$\tilde{S}(\vec{v}) = \arg\min_{\vec{S}\in\mathcal{M}}\vec{S}\cdot\vec{v}. \tag{54}$$

For a given data manifold $\mathcal{M}^\mu$ and known values of q, k, one can sample from the anchor point distribution using the mean-field theory (see Appendix A 14):

$$\tilde{S}(\vec{t}, t_0) = \frac{\vec{v}^* - \vec{t}}{\frac{ck}{1+ck}(v_0^* - t_0)}, \tag{55}$$

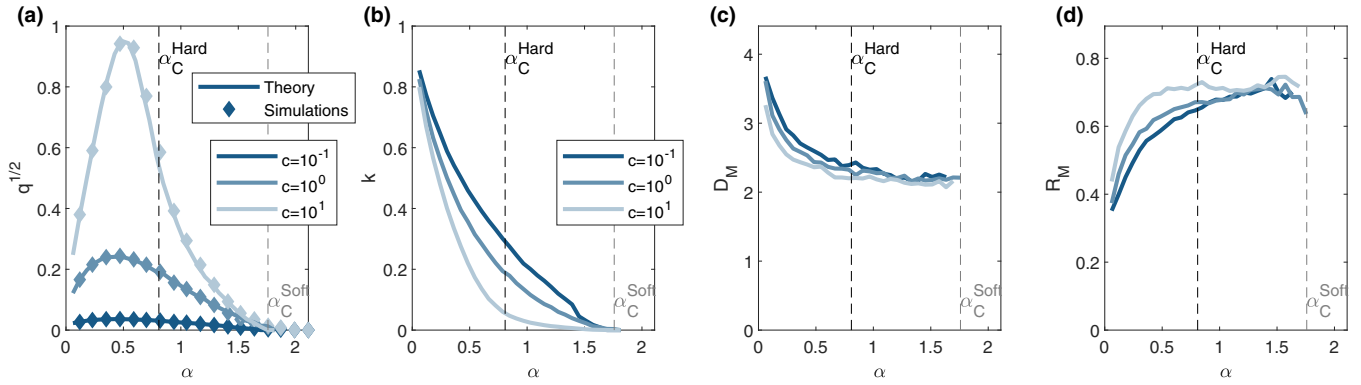

FIG. 10. Order parameters and manifold properties for point-cloud manifolds. Sampling $m = 100$ points from an ellipsoid with $\gamma = 1.5$, $R = 0.25, D = 20$. (a) The weights' norm $q^{1/2}$ (y axis) for different values of α (x axis) and choices of c (color coded). Compares theory (solid lines) and simulation results (diamonds). (b–d) The corresponding values of the order parameter k (b), manifold dimension D_M (c), and manifold radius R_M (d) (y axis) for different values of α (x axis) and choices of c (color coded).

where $\vec{v}^*, v_0^*$ are the values which minimize $F(\vec{t}, t_0)$, to be found using general least-squares optimization methods. This method for sampling from the anchor point distribution is formally described in Algorithm 4.

Large D regime We note that the structure of the manifold enters the mean-field equations only through the anchor points and their distribution. For large D, Ref. [7] has suggested their contribution can be summarized by measuring two statistics. Those manifold properties R_M, D_M are defined through the statistics of the anchor points with respect to $\vec{t}, t_0$:

$$R_M^2 = \langle \|\delta\tilde{S}\|^2 \rangle_{\vec{t},t_0}, \tag{56}$$

$$D_M = \langle (\vec{t} \cdot \delta\tilde{S})^2 / \|\delta\tilde{S}\|^2 \rangle_{\vec{t},t_0}. \tag{57}$$

As those generalize R, D of spheres, we suggest using R_M, D_M to solve for q, k, and α_C^{Soft} in the equations of soft classification of spheres. For each value of α, c we can iteratively calculate R_M, D_M by sampling anchor points using the current values of q, k, then update the estimation of q, k [using Eqs. (A36) and (A37) with $R = R_M, D = D_M$], until convergence (Algorithm 7). Similarly, we can calculate α_C^{Soft} directly by iteratively calculating R_M, D_M at small q, k, then update the estimation of α_C^{Soft} [using Eq. (41) with $R = R_M, D = D_M$], until convergence (Algorithm 8).

As was the case for spheres, when D is large we expect only the "touching" regime to contribute, and applying KKT condition to minimizing $F(\vec{t}, t_0)$ we get a self-consistent relation (see Appendix A 14):

$$\vec{v} = \vec{t} + \frac{ck}{1+ck}(1/\sqrt{q} - \vec{v} \cdot \tilde{S} - t_0)\tilde{S}. \tag{58}$$

Thus Eqs. (54) and (58) can be used to iteratively update $\vec{v}$ and $\tilde{S}$ (Algorithm 5). This iterative approach allows for finding the anchor points without solving a least-squares optimization problem for each value of $\vec{t}, t_0$, as in the least-squares algorithm.

Using manifold slacks we expect classification of general manifolds to exhibit finite capacity because the minimal field $h_{\min}$ is expected to be finite and negative relative to the central field v_0 as long as the weights are finite. To use a concrete example, for simulations of general manifolds we used point-cloud manifolds created by sampling m points from a D-dimensional ellipsoid with radii $r_l \sim l^{-\gamma}$. Denoting $R^2 = \sum_{l=1}^{D} r_l^2$ the ellipsoid shape is defined by parameters R, D, γ. Figures 10(a) and 10(b) and Figs. S16(a) and S16(b) demonstrate the existence of finite capacity when using manifold slacks also for those manifolds. The predicted values of q matches the empirically observed values, which vanish at a finite α value [Figs. 10(a) and S16(a)]. The dependence of the measured D_M on c and α is quite small [see Figs. 10(c) and S16(c)] and similarly for the measured R_M [see Figs. 10(d) and S16(d)].

Figure S18 presents the weights' norm for the classification of point-cloud manifolds and the theoretical values predicted for q, k, R_M, D_M, using either the iterative or the least-squares algorithm. The two algorithms give very similar results, with a notable difference at large R where the assumption that only the "touching" regime contributes to the solution no longer holds.

As it is favorable to have manifold properties D_M and R_M which do not depend on α, Fig. S19 shows that using a single choice of D_M, R_M, calculated for α near α_C^{Soft} (i.e., largest solvable α) to predict q provides a good match for the entire range of α (but not using a single choice calculated from a small α value).

Point-cloud manifolds are important for applications of the theory, and so we explored how manifold properties scale with the number of points in the manifold. Figure 11 demonstrates the contrast between structured manifolds, sampled from a low-dimensional ellipsoid, and random manifolds, sampled from Gaussian statistics. For both types the manifold radius R_M does not depend on the number of samples [Fig. 11(a)]. However, while for structured manifolds the manifold dimension D_M does not depend on the number of points [Fig. 11(b)], for random manifolds we observe that D_M is linear in the number of points m [Fig. 11(c)]. As a result, for structured manifolds the capacity α_C^{Soft} saturates to a finite value [Fig. 11(d)] when the number of samples is increased while for random manifolds it vanishes as $1/m$ [Fig. 11(e)]. The properties of random manifolds are further demonstrated in Fig. S20.

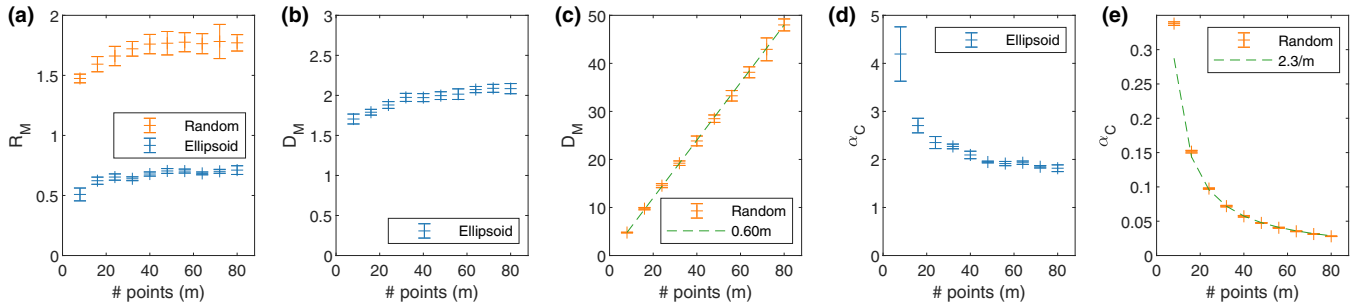

FIG. 11. Scaling of capacity and manifold geometry with the number of points in point-cloud manifolds. Comparison of point-cloud manifolds of m points sampled from either an ellipsoid with $\gamma = 1.5$, radius $R = 0.5$, and dimension $D = 10$, or from a random Gaussian. (a) Manifold radius (y axis) dependence on the number of manifold points (x axis) for ellipsoid and random manifolds. (b, c) Manifold dimension (y axis) dependence on the number of manifold points (x axis) for ellipsoid manifolds (b) and random manifolds (c). (d, e) Capacity (y axis) dependence on the number of manifold points (x axis) for ellipsoid manifolds (d) and random manifolds (e). All values calculated at the largest possible α; error bars indicate standard deviation with respect to choices of c.

Classification errors For general manifolds, the classification error is defined assuming manifold points are sampled according to some measure on the manifold (see discussion); for the simpler case of point-cloud manifolds, we assume this is a uniform distribution.

Figure S21 presents the training and test errors found in classification of point-cloud manifolds, demonstrating that the training error is monotonic in both c and α but the test error is not. Classification errors can be predicted from the theory of classification of spheres, by plugging in the theoretical values of q, k, R_M, D_M, calculated using the least-squares algorithm. Figure S22 compares the training error predicted with the error measured in simulations, and Fig. S23 compares the predicted test error at several noise levels with simulation results. The observed agreement means that the theory of spheres classification can be used to make predictions regarding classification of nonspherical manifolds by measuring the manifolds' R_M and D_M, demonstrates that those geometric properties capture the contribution of manifold shape to classification.

Comparison with other methods Comparing the performance of different classification methods on point-cloud manifolds reveals a similar behavior to that observed for spheres. Figure S24 compares the manifold-slack method with both center-slack and max-margin methods, using the optimal choice of c for each method. Below α_C^{Hard} manifold-slack classification exhibits improved performance compared to max-margin classification, but this improvement is usually small [Figs. S24(a) and S24(b)]. As in the case of spheres, for small α values manifold slacks are superior to center slacks, with large qualitative difference at low noise level when R is order 1, while for larger α values the performance of the center-slack method is better [Fig. S24(c)].

For point-cloud manifolds, when the number of samples per manifold is not too large, the point-slack method can also be used for manifold classification. Figures S25(a)–S25(d) show that using the point-slack method, there is no phase-transition to zero weights as for the manifold-slack method. Despite this marked difference, the classification error achieved by the point-slack method is only slightly better than that achieved by the manifold-slack method [both using the optimal choice of c; Figs. S25(e) and S25(f)]. This improvement is significant only at small levels of noise and towards α_C^{Soft}. Thus point-slack SVM uses the additional degrees of freedom (and additional computational costs) from assigning a separate slack variable per sample to slightly outperform the manifold-slack method.

III. DISCUSSION

The introduction of slack variables to SVMs allows linear classification of data which are not linearly separable, and for optimizing performance by choosing the right balance between making training errors and increasing classification margin [using the regularization parameter c; Eqs. (2), (23), and (25)]. Here we analyze the noise resilience of such classification by considering test performance with respect to input noise (with variance σ^2/N applied to each input component).

Point slack We first study the statistical mechanics of a point-slack model where a set of P random points in N dimensions are independently labeled, and each is assigned a slack variable. We show that the problem has a well defined solution for all load values $\alpha = P/N$ (Fig. 1). In the absence of input noise, the optimal choice of c is infinite for all α; however, in the presence of noise in the test data, the optimal c is finite (Fig. 2). Furthermore, the optimal choice of c can be calculated from theory [Eq. (18)], and is roughly given by the "canonical choice" $c = \sigma^{-2}$ (Fig. 3), demonstrating that an optimal regularization is tuned to the noise.

Manifold classification Our main interest is the case of points arranged in P randomly labeled manifolds, such that all points within a manifold have the same target label. Assuming the number of points per manifold is large (and possibly infinite) assigning a slack variable to each point is not feasible. We introduced and analyzed two schemes of slack algorithms for classification of manifolds, which differ in the manner in which slack variables are attached to manifolds. In the center-slack method, each manifold center is associated with a slack variable, reducing the learning to point-slack SVM of the centers. In the manifold-slack method, a slack variable is associated with the "worst" point in each manifold, relative to the separating hyperplane. The relation between slack variables and errors is different in the two methods (Fig. 4); when using center slacks, if the center is misclassified, most of the

manifold may follow, but using manifold slacks most of it may be classified correctly even if the "worst" point is not.

Center slack The relatively simple center-slack scheme has several attractive features. First, it has a well-defined, nonzero, solution for the weights for all values of α. Second, the associated optimal c is provided by theory (Fig. 5) and is approximately given by the simple "canonical choice" $c = (R^2 + \sigma^2)^{-1}$, where R is the manifold radius, expressing the intuition that the variability of the manifold data relative to the center (quantified by R^2) is an intrinsic noise on top of the extrinsic noise σ^2. Finally, for large α values its performance is superior to the more sophisticated manifold-slack method [Figs. 9(b) and 9(c) and Fig. S24(c)], as discussed below. The disadvantages of the center-slack method are its performance for small α values and that it does not generalize max-margin manifold classification.

Manifold slack The manifold-slack scheme is a natural extension of max-margin manifold classification [6,7] in which the optimal weight vector is a sum of anchor points, one per manifold, which are the closest points in each manifold to the separating hyperplane. Here each such point is assigned a slack variable. For α below the errorless classification capacity α_C^{Hard}, when c approaches ∞, manifold-slack classification approaches max-margin classification. However, the optimal c may not be infinite even in this α regime in the presence of noise (Fig. 8). As for larger values of α, a surprising result of our mean-field theory is that the manifold-slack method possesses a solution with nonzero weight vector only below a second critical value, α_C^{Soft} (Fig. 6). Thus, this method allows for extending the range of linear classification above the errorless capacity, but for a limited range (Fig. 7).

Using an optimal choice of c, the classification-error performance of manifold slacks is always better than max-margin and may be superior to center slacks, depending on parameters. The main improvement over max-margin is the extended range of α values (Fig. 7), as the reduction of the classification error is usually small [Figs. 9(a) and Figs. S13, S24(a), and S24(b)]. The improved performance compared to center slacks is substantial for small α values when the noise is small and R is order 1, where manifold slacks achieve near-zero error while center-slack error is order 1 [Figs. 9(b) and 9(c) and Fig. S24(c)].

While many of the results for manifolds were derived in the context of spheres, the theory extends well to general manifolds by recovering their effective radius and dimension [Eqs. (56) and (57); Figs. 10 and S16]. Importantly, their classification performance is well predicted by plugging those values into the theory of spheres (Figs. S22 and S23), thus demonstrating they capture the classification-relevant aspects of manifolds' geometry. For structured point-cloud manifolds we find a dimension which does not depend on the number of points, in sharp contrast to random point-cloud manifolds, where the dimension is linear in this number (Fig. 11).

Extension to other formalisms Soft classification can be defined in many ways [3,4,20], and each may be extended to manifold classification. Based on our analysis we expect that attaching a slack variable to the "worst" manifold point would lead to the same phase transition reported here. Figure S17 demonstrates this for a variant of manifold slack where an L_1 norm is used on the slacks, where the weights' norm vanishes at a load that is independent of c. On the other hand, attaching slack variables to predefined manifold locations would not lead to the appearance of such a transition. Furthermore, when attaching slack variables to both predefined points and the "worst" point (previously done in [21]), a phase transition is expected.

Measure on manifolds The use of manifold slacks benefits from being insensitive to the exact measure assumed on the manifolds (as long as it is nonzero). In the case of center slacks, the center of mass of the manifolds depends in general on the measure. Nevertheless, in some cases, there is a natural choice for the center, as in spheres or ellipsoids (due to symmetry), or in a points cloud, where using the points' average corresponds to a uniform measure on the points. Furthermore, one can use the measure-independent Steiner point [22] as the manifold center. Regardless of the employed classification method, the evaluation of the errors depends in general on the measure.

Future work Extending the theory of max-margin classification of manifolds to soft classification is an important step in connecting the theory to applications, where soft-margin classifiers are more commonly used. We believe the theory of general manifolds is relevant for the analysis of real-world data [23]. To properly do so, the theory needs to be extended to allow for center correlations, as was done for max-margin classifiers [8]; we expect this to be straightforward as the methods of [8] involve manifold preprocessing which is independent of the geometrical analysis.

The issue of robustness to noise would naturally come up when aiming to apply the theory to neural data analysis where noise is a common attribute of the problem, unlike the artificial networks analyzed in [8]. It would be interesting to apply the methods described here to analyze object representations with non-Gaussian noise, such as neural noise with Poisson-like characteristics.

On a broader scope, the discussion of robustness to noise is a limited form of generalization. In general, we would like to be able to discuss generalization with respect to a finite number of samples from a manifold, where the scaling behavior of the classification error with the number of samples is an open question. Recent work on the few-shot learning setup, where the number of samples is very small, has revealed relatively simple behavior of the classification error [24].

ACKNOWLEDGMENTS

H.S. is partially supported by the Gatsby Charitable Foundation, the Swartz Foundation, the National Institutes of Health (Grant No. 1U19NS104653), and the MAFAT Center for Deep Learning.

APPENDIX

1. Optimal loss in soft classification

We write a Lagrangian for the problems of points [Eq. (2)] and spheres [Eq. (25)], assuming no bias for brevity. For spheres the constraint on the minimal field is $h^{\mu}_{\mathrm{min}} = v^{\mu}_0 - R\|v^{\mu}\| \geqslant 1 - s^{\mu}$ so that both cases are captured by the

Lagrangian (with $R = 0$ for points)

$$\mathcal{L} = \|\boldsymbol{w}\|^2/N + c\|\vec{s}\|^2/N + 2\sum_{\mu}^{P}\beta_\mu\left(1 - s^\mu - h^\mu_{\min}\right). \quad \text{(A1)}$$

KKT conditions yield three equations, $0=\frac{\partial\mathcal{L}}{\partial w_i}$, $0 = \frac{\partial\mathcal{L}}{\partial s^\mu}$, and $0 = \vec{\beta}(1 - \vec{s} - \vec{h}_{\min})$. Those lead to $\vec{\beta} = \vec{s}c/N$ and $\|\boldsymbol{w}\|^2/N = \vec{\beta}^T\vec{h}_{\min}$, and hence the optimal solution $\mathcal{L}^*$ satisfies

$$\mathcal{L}^* = \sum_{\mu}^{P}\beta_\mu = \frac{c}{N}\sum_{\mu}^{P}s_\mu = c\alpha\langle s\rangle. \quad \text{(A2)}$$

2. Replica theory for points

Consider P points $\boldsymbol{x}^\mu \in \mathbb{R}^N$ and labels $y^\mu \in \{\pm 1\}$; soft-margin classification is defined as solving

$$\boldsymbol{w}^*, \vec{s}^* = \arg\min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 \text{ s.t. } h^\mu \geqslant 1 - s^\mu, \quad \text{(A3)}$$

where $\boldsymbol{w}^* \in \mathbb{R}^N$ and $\vec{s}^* \in \mathbb{R}^P_+$ and the fields $h^\mu = y^\mu(\boldsymbol{w}\cdot\boldsymbol{x}^\mu + b)$ where we assume $b = 0$ for brevity. Denoting the optimal loss as L^*, we write an expression for the volume of solutions $V(L, c)$, which vanishes for $L < L^*$:

$$V(L,c) = \int d^N\boldsymbol{w}\int d^P\vec{s}\prod_{\mu}^{P}\Theta(h^\mu - 1 + s^\mu)\delta(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL) \quad \text{(A4)}$$

$$= \int d^N\boldsymbol{w}\int d^P\vec{s}\int_{1-s^\mu}^{\infty}d^Ph^\mu\int\frac{d^P\hat{h}^\mu}{2\pi}e^{i\sum_\mu^P(y^\mu\boldsymbol{w}\cdot\boldsymbol{x}^\mu - h^\mu)\hat{h}^\mu}\int\frac{d\hat{l}}{2\pi}e^{i(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL)\hat{l}}. \quad \text{(A5)}$$

We wish to calculate the values for which the volume vanishes assuming random (Gaussian) points $\boldsymbol{x}^\mu$ and random (binary) labels y^μ. Using the replica identity [Eq. (5)] it is enough to find G which satisfies $[V^n] = e^{nG}$, to have that $[\log V] \approx G$. Thus we consider V^n, average over $x_i^\mu \sim \mathcal{N}(0, 1/N)$, and denote $q_{\alpha\beta} = \frac{1}{N}\sum_i^N w_i^\alpha w_i^\beta$. After integrating over $\hat{h}^{\alpha,\mu}$, w_i^α we have

$$[V^n]_x = \int d^{n\times n}q_{\alpha\beta}\int\frac{d^{n\times n}\hat{q}_{\alpha\beta}}{2\pi}\int\frac{d^n\hat{l}^\alpha}{\sqrt{2\pi}}e^{-nNG_0 - nNG_1}, \quad \text{(A6)}$$

$$G_0 = \frac{i}{n}\sum_{\alpha,\beta}^{n}q_{\alpha\beta}\hat{q}_{\alpha\beta} + \frac{1}{2n}\log\det(-2i\hat{q}_{\alpha\beta} - \delta_{\alpha\beta}2i\hat{l}^\alpha) + \frac{i}{n}\sum_{\alpha}^{n}L\hat{l}^\alpha, \quad \text{(A7)}$$

$$G_1 = \frac{\alpha}{2n}\log\det q - \frac{\alpha}{n}\log\int\frac{d^ns^\alpha}{\sqrt{2\pi}}\int_{1-s^\alpha}^{\infty}d^nh^\alpha e^{ic\sum_\alpha^n(s^\alpha)^2\hat{l}^\alpha - \frac{1}{2}\sum_{\alpha,\beta}^n q^{-1}_{\alpha\beta}h^\alpha h^\beta}. \quad \text{(A8)}$$

We assume replica symmetry, i.e., $q_{\alpha\beta} = q + (q_0 - q)\delta_{\alpha\beta}$, $-i\hat{q}_{\alpha\beta} = \hat{q} + (\hat{q}_0 - \hat{q})\delta_{\alpha\beta}$, and $-i\hat{l}^\alpha = \hat{l}$, and also that the behavior in the thermodynamic limit $N \to \infty$ is dominated by the maximum of the integral, so we can set $0 = \frac{\partial G_0}{\partial\hat{q}} = \frac{\partial G_0}{\partial\hat{q}_0}$ to get rid of those two variables:

$$G_0 = -\frac{1}{2} + (q_0 - L)\hat{l} - \frac{1}{2}\log(q_0 - q) - \frac{1}{2}\frac{q}{q_0 - q}. \quad \text{(A9)}$$

We use the Hubbard-Stratonovich transform to decouple G_1 into n terms, then use the replica identity $\log\int Dt\, z(t)^n \approx n\int Dt\log z(t)$ for $n \to 0$. Integrating over the slack variables s, denoting $k = 2\hat{l}(q_0 - q)$ and taking the limit $q \to q_0$, the integral in G_1 is dominated by the maximal value, given by $\frac{ckq_0}{1+ck}\min_{h>1/\sqrt{q}}(h - t)^2$. In the limit of $q \to q_0$:

$$\lim_{q\to q_0}(q_0 - q)G_0 = (q_0 - L)\frac{k}{2} - \frac{1}{2}q_0, \quad \text{(A10)}$$

$$\lim_{q\to q_0}(q_0 - q)G_1 = \frac{\alpha}{2}\frac{ckq_0}{1+ck}\alpha_0^{-1}(1/\sqrt{q_0}) \quad \text{(A11)}$$

for Gardner's $\alpha_0^{-1}(\kappa) = \int_{-\infty}^{\kappa}Dt(\kappa - t)^2$ [5]. Denoting $G = G_0 + G_1$, the volume vanishes when $\lim_{q\to q_0}(q_0 - q)G = 0$, yielding Eq. (6).

3. Self-consistent equations for points

The self-consistent equations for points, Eqs. (7) and (8), are derived directly from the mean-field Eq. (6) by assuming that for the optimal loss we expect saddle-point conditions on $L(q, k)$, namely, that $0 = \frac{\partial L}{\partial q} = \frac{\partial L}{\partial k}$.

Those self-consistent equations can be evaluated for the limits $\alpha \to 0$ and $\alpha \to \infty$. When $\alpha \to 0$ we have $k \to 1$ and $q \to 0$ so that $\alpha_0^{-1}(1/\sqrt{q}) \approx 1/q$, and thus $k \approx 1 - \frac{c}{1+c}\alpha$ while $\sqrt{q} \approx \frac{c}{1+c}\sqrt{\alpha}$. When $\alpha \to \infty$ we have $k \to 0$ and $q \to 0$ so that scaling $k = k_0/\alpha$ we have $k \approx 1/c\alpha$ and $q \approx 1/\alpha$.

The limit $c \to \infty$ exhibits different behavior for $\alpha < 2$ and $\alpha > 2$. For $\alpha < 2$ the problem follows the Lagrangian from Eq. (10), and the solution satisfies $1 = \alpha\alpha_0^{-1}(1/\sqrt{q})$, which is the max-margin solution. On the other hand, when $c \to \infty$ for $\alpha \geqslant 2$ the problem follows the Lagrangian from Eq. (11), and we have that $\lim_{c\to\infty}k = 0$ with finite q and $K = \lim_{c\to\infty}ck$, which obey the self-consistent equations

$$1 = \frac{K^2}{(1+K)^2}\alpha\alpha_0^{-1}(1/\sqrt{q}), \quad \text{(A12)}$$

$$1 = \frac{K}{1+K}\alpha H(-1/\sqrt{q}), \quad \text{(A13)}$$

and the relation between α and q becomes $\alpha = \alpha_0^{-1}(1/\sqrt{q})/H^2(-1/\sqrt{q})$, yielding Eq. (12).

4. Field and slack distribution for points

The replica theory yields, without integrating over the slack variables s and using the notation $k = 2\hat{l}(q_0 - q)$, that the limit $q \to q_0$ is given by an optimization problem with a

Lagrangian:

$$\mathcal{L} = \tfrac{1}{2}(h - t\sqrt{q})^2 + \tfrac{1}{2}cks^2 + \lambda(1 - h - s), \tag{A14}$$

where from KKT conditions the solution satisfies

$$0 = \lambda(1 - h - s), \tag{A15}$$
$$\lambda = h - t\sqrt{q}, \tag{A16}$$
$$\lambda = cks. \tag{A17}$$

In the "interior" regime $s = 0$, $h = \sqrt{q}t$; in the "touching" regime $s > 0$, $h = 1 - s = \sqrt{q}t + cks$, yielding

$$h = \begin{cases} \frac{ck}{1+ck} + \frac{\sqrt{q}}{1+ck}t_0 & -\infty \leqslant t_0 \leqslant 1/\sqrt{q} \\ \sqrt{q}t_0 & 1/\sqrt{q} \leqslant t_0 \end{cases}, \tag{A18}$$

which can be written equivalently as Eq. (13). The slack variables satisfy $s = \max\{1 - h, 0\}$ or explicitly

$$s = \begin{cases} \frac{1}{1+ck} - \frac{\sqrt{q}}{1+ck}t_0 & -\infty \leqslant t_0 \leqslant 1/\sqrt{q} \\ 0 & 1/\sqrt{q} \leqslant t_0 \end{cases}. \tag{A19}$$

Interestingly, the slack distribution allows deriving the self-consistent Eqs. (7) and (8) without saddle-point assumption (i.e., without taking derivatives of L). From the definition of L we have that $L = q + \alpha c\langle s^2\rangle$ while for the optimal loss $L^* = \alpha c\langle s\rangle$ (see Appendix A 1). Combining these equations with Eq. (6) yields Eqs. (7) and (8). Furthermore, from the expression for $\langle s^2\rangle$ and the self-consistent Eq. (7), $\alpha\langle s^2\rangle = \frac{q}{c^2k^2}$ [Eq. (9)].

5. Classification error for points

The training error $\varepsilon_{tr} = P(h < 0)$ has a contribution only from the "touching" regime of the field distribution [Eq. (13)], so that

$$\varepsilon_{tr} = H(ck/\sqrt{q}). \tag{A20}$$

When i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2/N)$ is applied to each input component, as the weights are independent of this noise, the fields are affected by i.i.d. noise $\mathcal{N}(0, \sigma^2 q)$, i.e., $h_\sigma = h + \sigma\sqrt{q}\eta$ when η is a standard Gaussian variable. The noisy field distribution can be written explicitly by convolving the field distribution with Gaussian; but for an analytic analysis of the error it is useful to write an expression for the error directly:

$$\varepsilon_g = P(h + \sigma\sqrt{q}\eta < 0) = \langle H(h/\sigma\sqrt{q})\rangle_h. \tag{A21}$$

Using the field distribution [Eq. (13)] and replacing $g_1 = h(1 + ck)/\sqrt{q} - ck/\sqrt{q}$ and $g_2 = h/\sqrt{q}$ we have

$$\varepsilon_g = \int_{-\infty}^{1/\sqrt{q}} Dg_1 H\left(\frac{g_1\sqrt{q} + ck}{\sigma\sqrt{q}(1 + ck)}\right) + \int_{1/\sqrt{q}}^{\infty} Dg_2 H(g_2/\sigma). \tag{A22}$$

Using identity 10,010.4 from [25] we get an expression which is approximated for $\sigma \ll 1/\sqrt{q}$ as $\varepsilon_g \approx H(\mathcal{S})$ for $\mathcal{S}$ from Eq. (16).

6. Optimal choice of c for points

We may optimize the SNR $\mathcal{S}$ with respect to c:

$$c^* = \arg\min_c \mathcal{S}^{-2}. \tag{A23}$$

Taking its derivative should satisfy $0 = \frac{\partial\mathcal{S}^{-2}}{\partial c}$, yielding an expression for $\frac{\partial q}{\partial c}$. On the other hand, starting from the self-consistent Eqs. (7) and (8) and taking the derivative with respect to c, using the identity $\frac{\partial}{\partial c}q\alpha_0^{-1}(1/\sqrt{q}) = H(-1/\sqrt{q})\frac{\partial q}{\partial c}$ we have a second expression for $\frac{\partial q}{\partial c}$. Combining these two equations yields an expression without $\frac{\partial q}{\partial c}$ or $\frac{\partial k}{\partial c}$. Using the self-consistent equations again to substitute $\alpha\alpha_0^{-1}(1/\sqrt{q})$ and $\alpha H(-1/\sqrt{q})$ we get that the optimal c satisfies Eq. (18), which needs to be solved self-consistently as k depends on c. Furthermore, for $\sigma = 0$ we have no solution with finite c and nonzero q, thus proving that ε_{tr} is monotonic in c for any α.

7. Replica theory for spheres

We write an expression for the volume $V(L, c)$ for $L = \|\boldsymbol{w}\|^2/N + c\|\vec{s}\|^2/N$ which vanishes for $L < L^*$:

$$V(L,c) = \int d^N\boldsymbol{w}\int d^P\vec{s}\int d^P\vec{h}\prod_\mu^P \delta\left(v_0^\mu - R\|\vec{v}^\mu\| - h^\mu\right)\Theta(h^\mu - 1 + s^\mu)\delta(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL) \tag{A24}$$

$$= \int d^N\boldsymbol{w}\int d^P\vec{s}\int_{1-s^\mu}^{\infty} d^Ph^\mu\int\frac{d^P\hat{h}^\mu}{2\pi}e^{i\sum_\mu^P(v_0^\mu - R\|\vec{v}^\mu\| - h^\mu)\hat{h}^\mu}\int\frac{d\hat{l}}{2\pi}e^{i(\|\boldsymbol{w}\|^2 + c\|\vec{s}\|^2 - NL)\hat{l}}, \tag{A25}$$

where we denote $v_l^\mu = y^\mu\boldsymbol{w}\cdot\boldsymbol{u}_l^\mu$ for $l = 0, \ldots, D$ and $\mu = 1, \ldots, P$ and enforce this using appropriate $\hat{v}_l^\mu$ variables.

We wish to calculate the values for which the volume vanishes assuming random (Gaussian) axes $\boldsymbol{u}_l^\mu$ and random (binary) labels y^μ. Using the replica identity [Eq. (5)] it is enough to find G which satisfies $[V^n] = e^{nG}$, to have that $[\log V] \approx G$. Thus we consider V^n and use a Gaussian integral on the axes $u_{li}^\mu \sim \mathcal{N}(0, 1/N)$, denoting as usual $q_{\alpha\beta} = \frac{1}{N}\sum_i^N w_i^\alpha w_i^\beta$. After Gaussian integration over $\hat{v}_l^{\alpha,\mu}$, w_i^α we have

$$[V^n]_x = \int d^{n\times n}q_{\alpha\beta}\int\frac{d^{n\times n}\hat{q}_{\alpha\beta}}{2\pi}\int\frac{d^n\hat{l}^\alpha}{\sqrt{2\pi}}e^{-nNG_0 - nNG_1}, \tag{A26}$$

$$G_0 = \frac{i}{n}\sum_{\alpha,\beta}^n q_{\alpha\beta}\hat{q}_{\alpha\beta} + \frac{1}{2n}\log\det(-2i\hat{q}_{\alpha\beta} - \delta_{\alpha\beta}2i\hat{l}^\alpha) + \frac{i}{n}\sum_\alpha^n L\hat{l}^\alpha, \tag{A27}$$

$$G_1 = \frac{\alpha}{2n}(D+1)\log\det q - \frac{\alpha}{n}\log\int\frac{d^n\vec{s}^\alpha}{\sqrt{2\pi}}\int d^{n\times D}\vec{v}_l^\alpha\int_{1-s^\alpha+R\|\vec{v}^\alpha\|}^{\infty}\frac{d^nv_0^\alpha}{\sqrt{2\pi}}e^{-\frac{1}{2}\sum_{l=0}^D\sum_{\alpha,\beta}^n q_{\alpha\beta}^{-1}v_l^\alpha v_l^\beta + i\sum_\alpha^n c(s^\alpha)^2\hat{l}^\alpha}. \tag{A28}$$

We assume replica symmetry, namely, $q_{\alpha\beta} = q + (q_0 - q)\delta_{\alpha\beta}$, $-i\hat{q}_{\alpha\beta} = \hat{q} + (\hat{q}_0 - \hat{q})\delta_{\alpha\beta}$, and $-i\hat{l}^\alpha = \hat{l}$ and also that the behavior in the thermodynamic limit $N \to \infty$ is dominated by the maximum of the integral, so that the derivatives satisfy $0 = \frac{\partial G_0}{\partial \hat{q}} = \frac{\partial G_0}{\partial \hat{q}_0}$. This allows for getting rid of those two variables, and G_0 becomes

$$G_0 = -\frac{1}{2} + (q_0 - L)\hat{l} - \frac{1}{2}\log(q_0 - q) - \frac{1}{2}\frac{q}{q_0 - q}. \quad \text{(A29)}$$

For G_1 we use the Hubbard-Stratonovich transform on $\sum_\alpha^n v_l^\alpha$ for $l = 0, \ldots, D$, so that G_1 decouples into n terms, and then use the replica identity $\log \int Dt\, z(t)^n \approx n \int Dt \log z(t)$ for $n \to 0$. We proceed by integrating away the slack parameters s and by completion to square of $v_l - t_l$. Renaming $k = 2\hat{l}(q_0 - q)$ and taking the limit $q \to q_0$, the integral is dominated by the maximum $F(\vec{t}, t_0; ck, q)$ from Eq. (33).

Taking the limit $q \to q_0$ yields

$$\lim_{q\to q_0}(q_0 - q)G_0 = (q_0 - L)\frac{k}{2} - \frac{1}{2}q_0, \quad \text{(A30)}$$

$$\lim_{q\to q_0}(q_0 - q)G_1 = \frac{\alpha q_0}{2}\int D^D\vec{t}\int Dt_0 F(\vec{t}, t_0; ck, q). \quad \text{(A31)}$$

Thus for $G = G_0 + G_1$ the volume vanishes at $\lim_{q\to q_0}(q_0 - q)G = 0$, yielding Eq. (32).

8. Solving the mean-field minimization problem for spheres

Let us solve the minimization problem from Eq. (33), so that we can write it as a closed-form expression. Denoting a Lagrangian,

$$\mathcal{L} = \frac{1}{2}\|\vec{v} - \vec{t}\|^2 + \frac{1}{2}\frac{ck}{1 + ck}(v_0 - t_0)^2 \cdots + \lambda(1/\sqrt{q} + R\|\vec{v}\| - v_0). \quad \text{(A32)}$$

From KKT conditions we have the equations

$$\lambda = \frac{ck}{1 + ck}(v_0 - t_0), \quad \text{(A33)}$$

$$t_l = v_l + \lambda R v_l / v, \quad \text{(A34)}$$

$$0 = \lambda(1/\sqrt{q} + Rv - v_0), \quad \text{(A35)}$$

denoting $v = \|\vec{v}\| \geqslant 0$ (and similarly we denote below $t = \|\vec{t}\|$). We solve those for different regimes:

(1) "Interior" regime, defined as $\lambda = 0$, where $v_0 = t_0$ and $v_l = t_l$ so that $F = 0$, valid at $t_0 \geqslant 1/\sqrt{q} + Rt$.

(2) "Embedded" regime, defined as $\lambda > 0$ and $v = 0$, such that $v_0 = 1/\sqrt{q}$ and $F = \frac{ck}{1+ck}(1/\sqrt{q} - t_0)^2 + t^2$.

(3) "Touching" regime, defined as $\lambda > 0$ and $v > 0$, with $v_0 = 1/\sqrt{q} + Rv$ and $v_l = \frac{v}{v+\lambda R}t_l$, which is valid at $1/\sqrt{q} - \frac{1+ck}{ck}t/R \leqslant t_0 \leqslant 1/\sqrt{q} + Rt$ and leads to $F = \frac{ck}{1+ck+ckR^2}(1/\sqrt{q} + Rt - t_0)^2$.

so that the minimization problem depends only on t_0 and the norm $t = \|\vec{t}\|$. As $t \sim \chi_D$ the chi distribution with D degrees of freedom, denoting $\chi_D(t) = \Gamma(D/2)^{-1}2^{1-D/2}t^{D-1}e^{-t^2/2}dt$ we have Eq. (34).

9. Self-consistent equations for spheres

Assuming the optimal loss satisfies the saddle-point equations $0 = \frac{\partial L}{\partial k} = \frac{\partial L}{\partial q}$ we have Eqs. (37) and (38). Taking the derivatives of f with respect to k, q yields the following self-consistent equations:

$$1 = \alpha\frac{(ck)^2(1 + R^2)}{(1 + ck(1 + R^2))^2}\int \chi_D(t)\int_{1/\sqrt{q}-\frac{1+ck}{ck}t/R}^{1/\sqrt{q}+Rt} \times Dt_0(1/\sqrt{q} + Rt - t_0)^2 + \alpha\int \chi_D(t)\int_{-\infty}^{1/\sqrt{q}-\frac{1+ck}{ck}t/R} Dt_0 \times \left[\frac{(ck)^2}{(1 + ck)^2}(1/\sqrt{q} - t_0)^2 + t^2\right], \quad \text{(A36)}$$

$$1 - k = \alpha\int \chi_D(t)\int_{1/\sqrt{q}-\frac{1+ck}{ck}t/R}^{1/\sqrt{q}+Rt} Dt_0\frac{ck}{1 + ck(1 + R^2)} \times (1/\sqrt{q} + Rt - t_0)(Rt - t_0) + \alpha\int \chi_D(t)\int_{-\infty}^{1/\sqrt{q}-\frac{1+ck}{ck}t/R} Dt_0 \times \left[t^2 - \frac{ck}{1 + ck}(1/\sqrt{q} - t_0)t_0\right]. \quad \text{(A37)}$$

As for points, those equations can also be derived by combining the equations for the optimal loss, namely, the loss definition $L = q + \alpha c\langle s^2\rangle$, the mean-field equation $L = q + \frac{q}{k}(\alpha f - 1)$ [Eq. (36)], and an optimality condition for the loss $L = c\alpha\langle s\rangle$ (see Appendix A 1), where the slack distribution [Eq. (A49)] leads to the self-consistent Eqs. (37) and (38) by noting the moments can be written as $\langle s^2\rangle = \frac{q}{c}\frac{\partial}{\partial k}f$ and $\langle s\rangle = -\frac{q}{ck}q\frac{\partial}{\partial q}f$.

10. Interesting regimes of the self-consistent equations for spheres

The self-consistent Eqs. (A36)–(A37) can be integrated over t_0 to yield equivalent consistent equations, which are useful when analyzing the behavior of different limits. Furthermore, when $D \gg 1$ the distribution of χ_D is narrow with a mode at $\sqrt{D-1}$ and a mean just below $\sqrt{D}$, so we may assume $t = \sqrt{D}$ and $\alpha_C \approx \frac{1+R^2}{R^2 D}$.

In the limit $c \to \infty$ for $\alpha < \alpha_C^{\text{Hard}}$ the problem follow the Lagrangian from Eq. (10) and converge with the max-margin case [7], while for $\alpha > \alpha_C^{\text{Hard}}$ the equations can be derived by a replica theory for the Lagrangian from Eq. (11). The resulting equations are related to the self-consistent equations of soft classification theory through $\lim_{c\to\infty} k = 0$ while q and $K = \lim_{c\to\infty} ck$ are finite, similarly to Eqs. (A12) and (A13) for points.

In the limit of $\alpha \to 0$ we expect to have $q \to 0$ and $k \to 1$, as in the case of soft classification of points, so that $1/\sqrt{q} - \frac{1+ck}{ck}\sqrt{D}/R \gg 1$ and $1/\sqrt{q} + R\sqrt{D} \gg 1$, and the self-consistent equations are simplified, resulting in the following first-order approximations for small α, $k \approx 1 - \alpha(1 + D)$ and $\sqrt{q} \approx \sqrt{\alpha}\frac{ck}{1+ck}$.

On the other hand, for $\alpha \to \alpha_C^{\text{Soft}}$ we expect both $q \to 0$ and $k \to 0$, so that we need to assume $1/\sqrt{q} + R\sqrt{D} \gg 1$

and $\frac{1+ck}{ck}\sqrt{D}/R - 1/\sqrt{q} \gg 1$, leading to different simplified equations, and the resulting order parameters $k \approx (1 - \sqrt{\alpha/\alpha_C})/(\alpha c + 1)$ for α_C from Eq. (43) and furthermore

$$\sqrt{q} \approx ck\frac{1+R^2}{R\sqrt{D}}. \tag{A38}$$

11. Capacity in classification of spheres

Consider the self-consistent Eqs. (A36)–(A37), and let us assume both k, $\sqrt{q} \ll 1$ and further that $k = x\sqrt{q}$. For the first equation we have two contributions $\int_0^{xcR} \chi_D(t)(t^2 + c^2x^2)$, and $c^2x^2(1+R^2)\int_{xcR}^{\infty}\chi_D(t)$, yielding

$$1 = \alpha\int_0^{xcR}\chi_D(t)(t^2+c^2x^2) + \alpha c^2x^2(1+R^2)\int_{xcR}^{\infty}\chi_D(t). \tag{A39}$$

For the second equation we two contributions $\int_0^{xcR}\chi_D(t)t^2$ and $xcR\int_{xcR}^{\infty}\chi_D(t)t$, leading to

$$1 = \alpha\int_0^{xcR}\chi_D(t)t^2 + \alpha xcR\int_{xcR}^{\infty}\chi_D(t)t. \tag{A40}$$

Combining these equations and replacing $xc \to x$ we have $x = kc/\sqrt{q}$, but the resulting equations are independent of c: a self-consistent Eq. (42) for x and Eq. (41) for $\alpha = \alpha_C$.

Now note that for $R \to 0$ we have that $x = R\int_{xR}^{\infty}\chi_D(t)t = R\sqrt{2}\Gamma(\frac{D}{2}+\frac{1}{2})/\Gamma(\frac{D}{2})$ and $\alpha_C^{-1} = x^2$ (which converges to R^2D for large D), whereas for $R \to \infty$ we have $x \approx 0$ and $\alpha_C^{-1} = D$. When $D \gg 1$ the distribution of χ_D is narrow around $\sqrt{D}$. If $\int_0^{xR}\chi_D \ll 1$ we have a much simpler result, as $x \approx \frac{R\sqrt{D}}{1+R^2}$ and thus

$$\alpha_C^{-1} = xR\sqrt{D} = \frac{R^2D}{1+R^2}. \tag{A41}$$

From the above limits on R we obtain that for large D this approximation is valid for any R.

12. Field and slack distribution for spheres

To derive the slack and field distribution we do not integrate away the slack variables and use the notation $k = 2\hat{l}(q_0 - q)$ to show that in the limit $\lim_{q_0\to q}(q_0-q)G_1$ the behavior is dominated by the solution of a constraint optimization problem with a Lagrangian:

$$\mathcal{L} = \frac{1}{2}\|\vec{v} - \sqrt{q}\vec{t}\|^2 + \frac{1}{2}(v_0 - \sqrt{q}t_0)^2\cdots + \frac{1}{2}cks^2 + \lambda(1 - s + R\|\vec{v}\| - v_0). \tag{A42}$$

and the solution should satisfy the KKT conditions:

$$0 = \lambda(1 - s + R\|\vec{v}\| - v_0), \tag{A43}$$

$$\lambda = v_0 - \sqrt{q}t_0, \tag{A44}$$

$$\sqrt{q}t_l = v_l(\|\vec{v}\| + \lambda R)/\|\vec{v}\|, \tag{A45}$$

$$\lambda = cks. \tag{A46}$$

Denoting $v = \|\vec{v}\|$ and $t = \|\vec{t}\|$ we have solution regimes:

(1) "Interior" regime: assuming $\lambda = 0$, which is valid for $t_0 \geqslant 1/\sqrt{q} + Rt$.

(2) "Touching" regime: assuming $\lambda > 0$, $v > 0$, which is valid for $1/\sqrt{q} - \frac{1+ck}{ck}t/R \leqslant t_0 \leqslant 1/\sqrt{q} + Rt$.

(3) "Embedded" regime: assuming $\lambda > 0$, $v = 0$ which is valid for $t_0 \leqslant 1/\sqrt{q} - \frac{1+ck}{ck}t/R$.

The fields and slack distribution can be written explicitly in terms of t, t_0 for the three regimes:

$$v_0 = \begin{cases} \frac{ck}{1+ck} + \frac{1}{1+ck}\sqrt{q}t_0 & \text{"Embedded"} \\ \frac{(1+R\sqrt{q}t)ck}{1+(1+R^2)ck} + \frac{(ckR^2+1)\sqrt{q}}{1+(1+R^2)ck}t_0 & \text{"Touching"} \\ \sqrt{q}t_0 & \text{"Interior"} \end{cases}, \tag{A47}$$

$$v = \begin{cases} 0 & \text{"Embedded"} \\ \frac{(1+ck)\sqrt{q}t - ckR}{1+(1+R^2)ck} + \frac{\sqrt{q}ckR}{1+(1+R^2)ck}t_0 & \text{"Touching"} \\ \sqrt{q}t & \text{"Interior"} \end{cases}, \tag{A48}$$

$$s = \begin{cases} \frac{1}{1+ck} - \frac{1}{1+ck}\sqrt{q}t_0 & \text{"Embedded"} \\ \frac{1+R\sqrt{q}t}{1+(1+R^2)ck} - \frac{\sqrt{q}}{1+(1+R^2)ck}t_0 & \text{"Touching"} \\ 0 & \text{"Interior"} \end{cases}, \tag{A49}$$

from which the slack variable moments, used for the self-consistent equations, are easily derived.

13. Classification error for spheres

Assuming $D \gg 1$, $t \sim \chi_D$ is concentrated around $\sqrt{D}$ and the distribution of v_0, v, s is a concatenation of the truncated Gaussian (or δ) distributions which correspond to the different regimes.

The slack distribution is then

$$s \sim \begin{cases} \mathcal{N}\left(\frac{1}{1+ck}, \frac{q}{(1+ck)^2}\right) & \frac{\sqrt{q}}{ck}\sqrt{D}/R < s \\ \mathcal{N}\left(\frac{1+R\sqrt{q}\sqrt{D}}{1+(1+R^2)ck}, \frac{q}{[1+(1+R^2)ck]^2}\right) & 0 < s \leqslant \frac{\sqrt{q}}{ck}\sqrt{D}/R \cdot \\ \delta(0)H\left(1/\sqrt{q} + R\sqrt{D}\right) & s = 0 \end{cases} \tag{A50}$$

Thus the probability of an error anywhere on the manifold is $\varepsilon_{tr}^{\text{manifold}} = P(s > 1) = H(\mathcal{S}^{\text{manifold}})$ for

$$\mathcal{S}^{\text{manifold}} = \begin{cases} ck/\sqrt{q} & \frac{\sqrt{q}}{ck}\sqrt{D}/R < 1 \\ (1+R^2)ck/\sqrt{q} - R\sqrt{D} & \frac{\sqrt{q}}{ck}\sqrt{D}/R \geqslant 1 \end{cases}, \tag{A51}$$

where for $\alpha \to \alpha_C^{\text{Soft}}$ we have that $(1+R^2)\sqrt{q}/ck = R\sqrt{D}$ so $\lim_{\alpha\to\alpha_C^{\text{Soft}}}\mathcal{S}^{\text{manifold}} = 0$.

Given any classifier $\boldsymbol{w}$, we assume the test error is calculated by sampling uniformly from the sphere, then adding noise. When i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2/N)$ is applied to each input component, as the weights are independent of this noise, the fields are affected by noise $\mathcal{N}(0, \sigma^2 q)$. That is, the error is given by $\varepsilon = P(h + \sigma\sqrt{q}\eta < 0) = \langle H(h/\sigma\sqrt{q})\rangle_h$ where η is a standard Gaussian variable.

For a D-dimensional spheres of radius R, denote the fields $h(\vec{S}) = y\boldsymbol{w}\cdot\boldsymbol{x}(\vec{S}) = v_0 + \vec{S}\cdot\vec{v}$. For a given $\boldsymbol{w}$, we can always choose the coordinate system such as $u_1 \propto w$ so that $v_1 = w\cdot u_1$ and $v_i = 0$ for $i > 1$, so that $v = \|\vec{v}\| = v_1$. Denote $S_1 = z$ we note that $\vec{S}\cdot\vec{v} = zv$ and thus $h = v_0 + zv$. As the joint distribution of v_0, v is given by theory [Eqs. (A47) and (A48)] it is enough to find the distribution of z under

uniform sampling from the sphere. As $z \in [-R, R]$, we can denote $\boldsymbol{x} \in \mathcal{S}_{D-2}(\sqrt{R^2 - z^2})$ a sphere of all choices for the values of $S_{2,\ldots,D}$; using the n-ball surface formula, $S_{n-1}(r) = 2\pi^{\frac{n}{2}}\Gamma(\frac{n}{2})^{-1}r^{n-1}$, the surface of the $D-1$ sphere with a radius $\|x\|$ is $S_{D-2}(\sqrt{R^2 - z^2})$, which needs to be normalized by the total surface, given by $S_{D-1}(R)$. Using polar coordinates the measure on z is given by $R/\sqrt{R^2 - z^2}$, and by a change of variable $\hat{z} = z/R$ we use the surface formulas to derive the bell-shaped distribution of $\hat{z}$, Eq. (45), supported at $\hat{z} \in [-1, 1]$. Then the error is an average with respect to $P(\hat{z})$, namely, $\varepsilon = \langle H[(v_0 + R\hat{z}v)/\sigma\sqrt{q}]\rangle_{\hat{z},v_0,v}$.

For $D \gg 1$, by assuming that only the "touching" regime contributes to the error, we may evaluate the leading orders of $v_0 + R\hat{z}v$ to derive a simpler expression for the error. The values of v_0, v in this regime are given by Eqs. (A47) and (A48) and depend on $t_0 \sim \mathcal{N}(0, 1)$ and $t \sim \chi_D$. Noting that $\hat{z}, t, t_0$ are pairwise independent, we can calculate the first two moments; then approximating $v_0 + R\hat{z}v$ as Gaussian and using $\langle H(x/a)\rangle_{x\sim\mathcal{N}(\mu,s^2)} = H(\mu/\sqrt{s^2 + a^2})$ we have the following approximation, denoting σ_0^2 the total contribution of the different terms to the variance:

$$\varepsilon \approx \left\langle H\left(\frac{(1/\sqrt{q} + Rt)ck}{\sqrt{\sigma_0^2 + [1 + (1 + R^2)ck]^2\sigma^2}}\right)\right\rangle_{t\sim\chi_D}, \tag{A52}$$

$$\sigma_0^2(t) \doteq \left(ckR^2 + 1\right)^2 + \frac{R^4}{D}(ck)^2 \times \left[\left(1/\sqrt{q} - \frac{1 + ck}{ck}t/R\right)^2 + 1\right], \tag{A53}$$

and the training error is given by setting $\sigma = 0$. Near α_C we have $k \to 0$ such that $\sigma_0^2 = 1 + 1/(R^{-1} + R)^2 \approx 1$ and using Eq. (A38) yields $\varepsilon \approx H(\frac{R\sqrt{D}}{1+R^2}\frac{1}{\sqrt{1+\sigma^2}})$.

14. Iterative algorithm for general manifolds

From the mean-field equations of spheres we get that a theory of general manifolds implies the same equation:

$$1 = (1 - L/q)k + \alpha \int D^D\vec{t} \int Dt_0 F(\vec{t}, t_0), \tag{A54}$$

where the inner minimization is now defined as

$$F(\vec{t}, t_0) = \min_{v_0 + g(\vec{v}) \geq 1/\sqrt{q}} \left\{\|\vec{v} - \vec{t}\|^2 + \frac{ck}{1 + ck}(v_0 - t_0)^2\right\}, \tag{A55}$$

where $g(\vec{v}) = \min_{\vec{S}\in M} \vec{v}\cdot\vec{S}$ is a scalar function with a subgradient $\tilde{S}(\vec{v}) = \frac{\partial}{\partial v}g(\vec{v})$. Denoting a Lagrangian

$$\mathcal{L} = \frac{1}{2}\|\vec{v} - \vec{t}\|^2 + \frac{1}{2}\frac{ck}{1 + ck}(v_0 - t_0)^2 + \lambda[1/\sqrt{q} - v_0 - g(\vec{v})], \tag{A56}$$

the optimal solution satisfies KKT conditions:

$$0 = \lambda[1/\sqrt{q} - v_0 - g(\vec{v})], \tag{A57}$$

$$\lambda = \frac{ck}{1 + ck}(v_0 - t_0), \tag{A58}$$

$$\vec{v} = \vec{t} + \lambda\tilde{S}(\vec{v}), \tag{A59}$$

so that we get Eq. (55) for $\tilde{S}(\vec{t}, t_0)$ when $v_0 \neq t_0$. Denoting $v = \|\vec{v}\|$ and $t = \|\vec{t}\|$ we have the following regimes:

(1) "Interior" regime: assuming $v > 0$ and $\lambda = 0$ we have $v_0 = t_0$ and $v_l = t_l$, so that $F = 0$.

(2) "Embedded" regime: assuming $v = 0$ and $\lambda > 0$ we have $v_0 = 1/\sqrt{q}$, so that $F = t^2 + \frac{ck}{1+ck}(1/\sqrt{q} - t_0)^2$.

(3) "Touching" regime: assuming $v > 0$ and $\lambda > 0$ we have $v_0 = 1/\sqrt{q} - g(\vec{v})$ and $\vec{v} = \vec{t} + \frac{ck}{1+ck}(1/\sqrt{q} - \vec{v}\cdot\tilde{S} - t_0)\tilde{S}$ [i.e., Eq. (58)], so that $F = \frac{ck}{1+ck(1+\tilde{S}^2)}(1/\sqrt{q} - t_0 - \vec{t}\cdot\tilde{S})^2$.

[1] V. Vapnik and A. Y. Lerner, Recognition of patterns with help of generalized portraits, Avtomat. i Telemekh **24**, 774 (1963).

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik, A training algorithm for optimal margin classifiers, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Association for Computing Machinery, New York, NY, 1992), pp. 144–152

[3] C. Cortes and V. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).

[4] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, New support vector algorithms, Neural Comput. **12**, 1207 (2000).

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

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

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

[8] U. Cohen, S. Chung, D. D. Lee, and H. Sompolinsky, Separability and geometry of object manifolds in deep neural networks, Nat. Commun. **11**, 746 (2020).

[9] D.-R. Chen, Q. Wu, Y. Ying, and D.-X. Zhou, Support vector machine soft margin classifiers: Error analysis, J. Mach. Learn. Res. **5**, 1143 (2004).

[10] J. Shawe-Taylor and N. Cristianini, On the generalization of soft margin algorithms, IEEE Trans. Inf. Theory **48**, 2721 (2002).

[11] C. Park, Convergence rates of generalization errors for margin-based classification, J. Stat. Plan. Inference **139**, 2543 (2009).

[12] R. Dietrich, M. Opper, and H. Sompolinsky, Statistical Mechanics of Support Vector Networks, Phys. Rev. Lett. **82**, 2975 (1999).

[13] S. Risau-Gusman and M. B. Gordon, Statistical mechanics of learning with soft margin classifiers, Phys. Rev. E **64**, 031907 (2001).

[14] S. Risau-Gusman and M. B. Gordon, Learning curves for soft margin classifiers, arXiv:cond-mat/0203315 (2002).

[15] A. Kammoun and M.-S. Alouini, On the precise error analysis of support vector machines, IEEE Open J. Signal Process. **2**, 99 (2021).

[16] M. Mézard, G. Parisi, and R. Zecchina, Analytic and algorithmic solution of random satisfiability problems, Science **297**, 812 (2002).

[17] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, World Scientific Lecture Notes in Physics, Vol. 9 (World Scientific, Singapore, 1987).

[18] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.106.024126 for additional figures and formal description of the proposed algorithms.

[19] H. W. Kuhn and A. W. Tucker, Nonlinear programming, in *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistics* (University of California Press, Berkeley, 1951), pp. 481–492.

[20] Q. Wu and D.-X. Zhou, Svm soft margin classifiers: Linear programming versus quadratic programming, Neural Comput. **17**, 1160 (2005).

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

[22] G. C. Shephard, The Steiner point of a convex polytope, Can. J. Math. **18**, 1294 (1966).

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

[24] B. Sorscher, S. Ganguli, and H. Sompolinsky, The geometry of concept learning, bioRxiv:10.1101/2021.03.21.436284 (2021).

[25] D. B. Owen, A table of normal integrals: A table, Commun. Stat.-Simul. Comput. **9**, 389 (1980).