PHYSICAL REVIEW RESEARCH 2, 023169 (2020)

Editors' Suggestion

Counting the learnable functions of geometrically structured data

Pietro Rotondo,[1,2] Marco Cosentino Lagomarsino,[3,4] and Marco Gherardi[3,4,*]

[1]*School of Physics and Astronomy, University of Nottingham, Nottingham NG7 2RD, England, United Kingdom*
[2]*Centre for the Mathematics and Theoretical Physics of Quantum Non-Equilibrium Systems, University of Nottingham, Nottingham NG7 2RD, England, United Kingdom*
[3]*Università degli Studi di Milano, Via Celoria 16, 20133 Milano, Italy*
[4]*Istituto Nazionale di Fisica Nucleare, Sezione di Milano, Milano, Italy*

(Received 5 November 2019; accepted 8 April 2020; published 13 May 2020)

Cover's function counting theorem is a milestone in the theory of artificial neural networks. It provides an answer to the fundamental question of determining how many binary assignments (dichotomies) of p points in n dimensions can be linearly realized. Regrettably, it has proved hard to extend the same approach to more advanced problems than the classification of points. In particular, an emerging necessity is to find methods to deal with geometrically structured data, and specifically with non-point-like patterns. A prominent case is that of invariant recognition, whereby identification of a stimulus is insensitive to irrelevant transformations on the inputs (such as rotations or changes in perspective in an image). An object is thus represented by an extended perceptual manifold, consisting of inputs that are classified similarly. Here, we develop a function counting theory for structured data of this kind, by extending Cover's combinatorial technique, and we derive analytical expressions for the average number of dichotomies of generically correlated sets of patterns. As an application, we obtain a closed formula for the capacity of a binary classifier trained to distinguish general polytopes of any dimension. These results extend our theoretical understanding of the role of data structure in machine learning, and provide useful quantitative tools for the analysis of generalization, feature extraction, and invariant object recognition by neural networks.

DOI: 10.1103/PhysRevResearch.2.023169

I. INTRODUCTION

Machine learning and deep learning demonstrate astonishing results in applications, sometimes beyond our theoretical reach. This provides a formidable challenge for theorists who wish to develop a framework for their understanding [1–8]. A landmark achievement in learning theory is Cover's function counting theorem, which counts the number of binary classification functions, or "dichotomies," that can be realized by given architectures [9]. This foundational result allowed theorists to quantify the complexity of a learning model and the advantage gained in using nonlinear kernels, provided a benchmark for the performance of both artificial and natural neural networks, and is a handy tool for several applications [10–15].

Other commonly used methods in this area come from statistical physics (pioneered by Gardner [16] and Gardner and Derrida [17]; see [18,19] for recent examples). With respect to these, Cover's method has the advantage of offering a simple geometric insight and of being valid at a finite number of dimensions, while statistical physics methods typically apply in the "thermodynamic limit" of infinite dimensions. Yet, despite its benefits and relative simplicity, Cover's analytical technique has so far eluded efforts to extend it [11].

Uncorrelated random patterns are commonly taken as a simplifying assumption for the theoretical investigation of artificial neural networks. Yet, it is becoming apparent that providing a theoretical framework that includes geometrical structure in the input data is essential. This need is emerging in different contexts.

(a) The invariant representation of perceptual stimuli by brains (e.g., the coherent perception of differently rotated and rescaled objects in vision, or the recognition of the same sound in different acoustic environments in audition) prompted the formalization of perceptual manifolds as geometrically extended patterns [15,20–27]. Perceptual manifolds are the regions in input space corresponding to all variations of a stimulus that do not modify the object's identification.

(b) The discovery of spatial maps in rodent brains [28] motivated extensions of associative memory models to attractors that are not pointlike but occupy a region in configuration space [29].

(c) The problem of local generalization and robustness to noise, a main theme of machine learning, can be cast as a problem of non-point-like patterns [30–32].

(d) The description of the input patterns as modular combinations of elementary features (a well-studied aspect of empirical datasets [33,34]) was shown to induce a multilayer structure in certain network architectures [35].

(e) Various properties of multilayer networks, related both to learning and to generalization, were observed to be strongly dependent on data structure [36,37].

Here, we develop a theory that extends Cover's approach to non-point-like patterns, by counting only those dichotomies

*Corresponding author: marco.gherardi@mi.infn.it

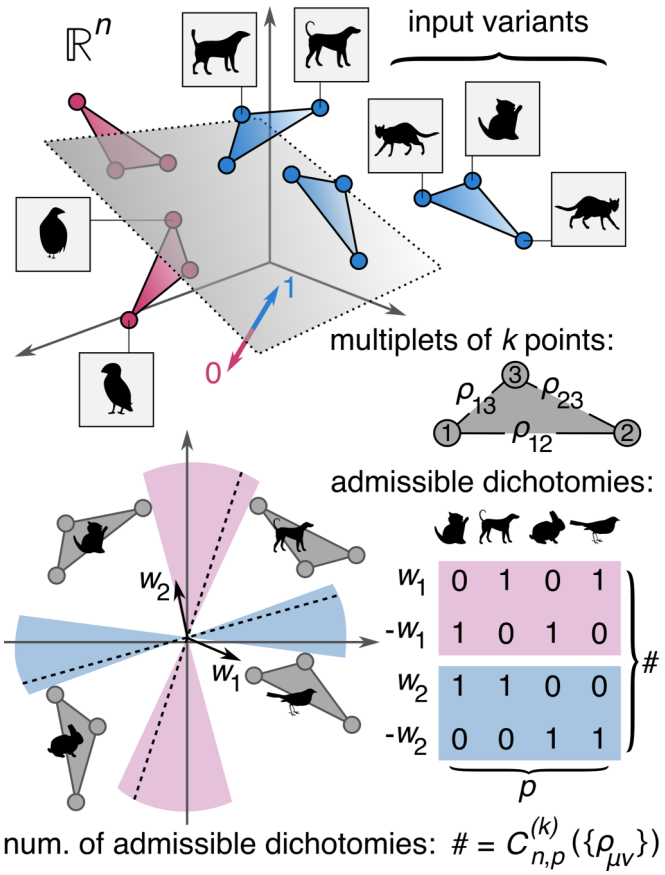

FIG. 1. Top: A dichotomy is identified by a hyperplane (in gray), separating differently labeled data points (e.g., mammals from birds, mapped, respectively, to 1 and 0). Data are structured in multiplets of k input variants (here $k = 3$). Each input variant is a point in $\mathbb{R}^n$, and each multiplet is characterized by the $k(k-1)/2$ overlaps between its points (here ρ_{12}, ρ_{23}, and ρ_{13}). A dichotomy is admissible only if it is constant on each multiplet, i.e., if the separating hyperplane does not intersect any polytope (triangles here). Bottom: Given a structured dataset (here $p = 4$ multiplets of $k = 3$ points in $n = 2$ dimensions), we count the number $C^{(k)}_{n,p}$ of admissible dichotomies.

that assign the same label to different variants of the same input. How variants within the same group are related to each other defines what we call the geometrical structure of the data. Our theory (i) enables the exact computation of the (average) number of dichotomies of structured data; (ii) gives direct access to quantities at finite size; (iii) naturally disentangles combinatorial and geometric aspects, thus lending itself to further generalizations; and (iv) is parametrized in terms of quantities easily measurable directly from data. These results are made possible by a hybrid approach, combining Cover's rigorous technique with mean-field approximations, inspired by statistical physics.

II. NUMBER OF ADMISSIBLE DICHOTOMIES

The central quantity obtained by Cover's function counting method is the number $C_{n,p}$ of linearly realizable dichotomies of p points $\xi_1, \ldots, \xi_p$ in n dimensions. A dichotomy of this set is a function ϕ mapping each point ξ_i to its $\{0, 1\}$ binary label (see Fig. 1). A linearly realizable dichotomy is identified by a vector $w \in \mathbb{R}^n$:

$$\phi(\xi_i) = \theta(\xi_i \cdot w), \tag{1}$$

where θ is the Heaviside theta function. The hyperplane perpendicular to the vector w separates the space into two half spaces, where the points mapped to 0 and 1 lie, respectively. There are 2^p dichotomies, but only $C_{n,p}$ of them are linearly realizable. We focus on linearly realizable dichotomies, and will therefore omit this specification when it is clear from the context.

It turns out that $C_{n,p}$ does not depend on the ξ_i's, as long as they are in general position (meaning that no subset of n or less points is linearly dependent) [9]. Structure in the data may thus appear not to affect $C_{n,p}$ at all. However, in general we do not wish to admit all possible dichotomies. For instance, among the hand-written digits in the popular MNIST (Modified National Institute of Standards and Technology) database we could choose to admit dichotomies separating "1" and "I," but not two similar-looking "0"s. Our definition of structure is based on such a restriction: a dataset is qualified as structured whenever only a subset of all possible dichotomies is considered admissible. $C_{n,p}$ will then be the number of admissible dichotomies that can be realized linearly. (Notice that this definition of geometrical structure is not related to the data being curated or possessing categorical features [38].)

Here we focus on a rather general definition of admissibility, inspired by the literature cited above. We consider datasets of kp points, structured as p multiplets of k points each. A dichotomy ϕ is admissible if different points ξ in the same multiplet are classified coherently, i.e., if $\phi(\xi)$ is constant on each multiplet. We will restrict the points ξ to lie on the unit sphere S^{n-1}, meaning that $\xi^2 = 1$, but this technical requirement can be easily relaxed. (A useful consequence of this is that setting the overlap between two points determines their distance.) The ensemble we consider fixes all the overlaps between the points in a multiplet, equally for all multiplets, but the relative positions and orientations of the multiplets are unspecified. The quantities we will compute are averages over all possible positions and orientations of the multiplets. More precisely, the (marginal) probability distribution for each point ξ_i is the flat distribution on the unit sphere S^{n-1}; the joint probability distribution $p(\{\xi_i\})$ is the product of the single point probabilities, conditioned to the constraint $\mathcal{C}$ that all overlaps within a multiplet are fixed [see Eqs. (5) and (22) below]:

$$p(\{\xi_i\}) = \int \prod_{i=1}^{kp} \mathcal{D}\xi_i \; \delta_{\mathcal{C}} \tag{2}$$

where $\mathcal{D}$ is the flat measure on S^{n-1} and $\delta_{\mathcal{C}}$ is the Dirac delta of the constraints.

Because of the convexity of linear separability, separating the multiplets is equivalent to separating the polytopes the vertices of which are the points in the multiplets. (These polytopes play the role of the perceptual manifolds of [15].) For instance, $k = 2$ corresponds to segments, $k = 3$ corresponds to triangles, and $k = 4$ corresponds to tetrahedra.

III. SINGLE POINTS ($k = 1$)

Let us first outline Cover's original computation. Imagine starting with p points and adding the $(p+1)$th point ξ_{p+1} to

$\{\xi_1, \ldots, \xi_p\}$. For each dichotomy ϕ of the p points $\xi_1, \ldots \xi_p$ one of two possibilities is satisfied: either (i) ϕ can be realized by a hyperplane passing through ξ_{p+1} (equivalently, ϕ can be realized by a vector w such that $\xi_{p+1} \cdot w = 0$) or (ii) it cannot. If (i) is true, then w can be rotated infinitesimally to yield both $\xi_{p+1} \cdot w \gtrless 0$; otherwise, the half space where ξ_{p+1} lies is fixed. Therefore, for each dichotomy ϕ of $\{\xi_1, \ldots, \xi_p\}$ satisfying (i) there are two different dichotomies ϕ_1 and ϕ_2 of $\{\xi_1, \ldots, \xi_p, \xi_{p+1}\}$ agreeing with ϕ on the common points [i.e., such that $\phi_{1,2}(\xi_i) = \phi(\xi_i)$ for $i = 1, \ldots, p$]. If the number of dichotomies satisfying (i) is M, then the number of those satisfying (ii) is $C_{n,p} - M$, and one can write $C_{n,p+1} = 2M + C_{n,p} - M$. The condition (i) is in the form of a single linear constraint, therefore M is the number of dichotomies of p points in $n-1$ dimensions, $M = C_{n-1,p}$. Thus $C_{n,p}$ satisfies the recursion

$$C_{n,p+1} = C_{n,p} + C_{n-1,p}, \tag{3}$$

with boundary conditions $C_{n>0,1} = 2$ (a single point can be classified either way) and $C_{0,p} = 0$.

The solution to Eq. (3) can be obtained by observing that the contribution of the boundary value $C_{n-i,1}$ to $C_{n,p}$ is given by the number of directed paths $\{\gamma_j\}_{j=1,\ldots,p}$, with $\gamma_j \in \mathbb{N}$, that start from $\gamma_1 = n - i$ and end in $\gamma_p = n$, where at each step γ_{j+1} can be either γ_j or $\gamma_j + 1$. The number of such paths is simply the binomial coefficient $\binom{p-1}{i}$. Summing over the boundary gives

$$C_{n,p} = 2\sum_{i=0}^{n-1} \binom{p-1}{i}, \tag{4}$$

where it is assumed that $\binom{p-1}{i} = 0$ whenever $i > p - 1$.

Let us consider the fraction $c_{n,p}$ of linearly realizable dichotomies $c_{n,p} = C_{n,p}/2^p$. For finite n and p, the capacity α_c can be defined as the ratio p/n at which half of all dichotomies can be realized: $c_{n,n\alpha_c} = 1/2$. From the explicit expression (4) one sees that $c_{n,p} = 1$ if $p \leqslant n$, $c_{n,p} \to 0$ for $p \to \infty$, and $c_{n,2n} = 1/2$, which pinpoints the well-known capacity $\alpha_c = 2$.

IV. SEGMENTS (DOUBLETS, $k = 2$)

The first step towards the general problem is the case where data are structured as pairs of points. Alongside the set of points $\xi = \{\xi_1, \ldots, \xi_p\}$, let us consider another set $\bar{\xi} = \{\bar{\xi}_1, \ldots, \bar{\xi}_p\}$. The multiplets discussed above are the doublets $\{\xi_i, \bar{\xi}_i\}$. Each doublet is such that the overlap between the two partners is fixed:

$$(-1, 1) \ni \rho = \xi_i \cdot \bar{\xi}_i \tag{5}$$

for all i. The admissible dichotomies ϕ are those for which $\phi(\xi_i) = \phi(\bar{\xi}_i)$ for all i; their total number is 2^p.

The recursion step now corresponds to the addition of the $(p+1)$th doublet $\{\xi_{p+1}, \bar{\xi}_{p+1}\}$. Repeating Cover's reasoning for the point $\bar{\xi}_{p+1}$ alone gives a number of dichotomies equal to $Q_{n,p} = C_{n,p} + C_{n-1,p}$. This is the number of dichotomies of the set $\{\xi_1, \bar{\xi}_1, \xi_2, \bar{\xi}_2, \ldots, \xi_p, \bar{\xi}_p, \bar{\xi}_{p+1}\}$ that are admissible on the first p doublets [meaning that $\phi(\xi_i) = \phi(\bar{\xi}_i)$ for all $i = 1, \ldots, p$]. A number $R_{n,p}$ of such dichotomies are realizable by a hyperplane passing through the point ξ_{p+1}. These are all admissible, thanks to the freedom in the choice of $\phi(\xi_{p+1})$ by an infinitesimal adjustment of the hyperplane. Among the other $Q_{n,p} - R_{n,p}$ dichotomies, on average, a fraction Ψ_2 will happen to assign the same label to ξ_{p+1} and $\bar{\xi}_{p+1}$. Ψ_2 can be computed as the fraction of hyperplanes keeping ξ_{p+1} and $\bar{\xi}_{p+1}$ in the same half space; the calculation is carried out in the Appendix. Importantly, Ψ_2 is a function of the overlap ρ alone:

$$\Psi_2(\rho) = \frac{2}{\pi} \arctan \sqrt{\frac{1+\rho}{1-\rho}}. \tag{6}$$

Note that $\Psi_2(\rho) = 1 - \Psi_2(-\rho)$ as expected from its definition. The foregoing argument brings us to estimate the total number of admissible dichotomies as

$$C_{n,p+1} = \Psi_2(\rho)(C_{n,p} + C_{n-1,p}) + [1 - \Psi_2(\rho)]R_{n,p}. \tag{7}$$

In order to compute $R_{n,p}$ it suffices to repeat Cover's reasoning with respect to the point $\bar{\xi}_{p+1}$, this time in $n-1$ dimensions because of the constraint imposed by the hyperplane passing through ξ_{p+1}, thereby obtaining

$$R_{n,p} = C_{n-1,p} + C_{n-2,p}. \tag{8}$$

Finally the recursion for $C_{n,p}$ reads

$$C_{n,p+1} = \Psi_2(\rho)C_{n,p} + C_{n-1,p} + [1 - \Psi_2(\rho)]C_{n-2,p}. \tag{9}$$

The boundary conditions are now slightly different than those for the case $k = 1$ in Eq. (3). In fact, in $n = 1$ dimension the number of admissible dichotomies of a single pair of points ($p = 1$) is 2 only when both points lie on the same half line, otherwise it is zero; on average, it is $2\Psi_2(\rho)$. The boundary conditions are then

$$C_{0,p} = 0, \quad C_{n>0,1} = 2\{1 - [1 - \Psi_2(\rho)]\delta_{n,1}\}. \tag{10}$$

To find the solution of the recursion (9), similarly to the single point case, consider all the directed paths $\{\gamma_j\}_{j=1,\ldots,p}$ propagating from the boundary to $C_{n,p}$, where γ_{j+1} at each step can be γ_j, $\gamma_j + 1$, or $\gamma_j + 2$. Contrary to the one point case, different paths with the same end points can now give different contributions to $C_{n,p}$, since the three types of steps correspond to three different factors (Ψ_2, 1, and $1 - \Psi_2$, respectively). The contribution $K_{i,p}$ of a path from $\gamma_1 = n - i$ to $\gamma_p = n$ is

$$K_{i,p} = \sum_{m=0}^{p-1} \binom{p-1}{m, i-2m} \Psi_2(\rho)^{p-1-i+m}[1 - \Psi_2(\rho)]^m, \tag{11}$$

where the multinomial coefficient is defined as

$$\binom{n}{m_1, m_2} = \frac{n!}{m_1! m_2! (n - m_1 - m_2)!} \tag{12}$$

(with the obvious analytical extension for negative factorials). Summation over the nonzero boundary $i = 0, \ldots, n-1$ yields the number of admissible dichotomies:

$$C_{n,p} = 2\sum_{i=0}^{n-2} K_{i,p} + 2\Psi_2(\rho)K_{n-1,p}. \tag{13}$$

It is easy to see (by the multinomial theorem) that $C_{n,p} = 2^p$ if $p \leqslant n/2$; this locates the usual Vapnik-Chervonenkis dimension [39], $d_{VC} = n$, as the total number of points is $2p$.

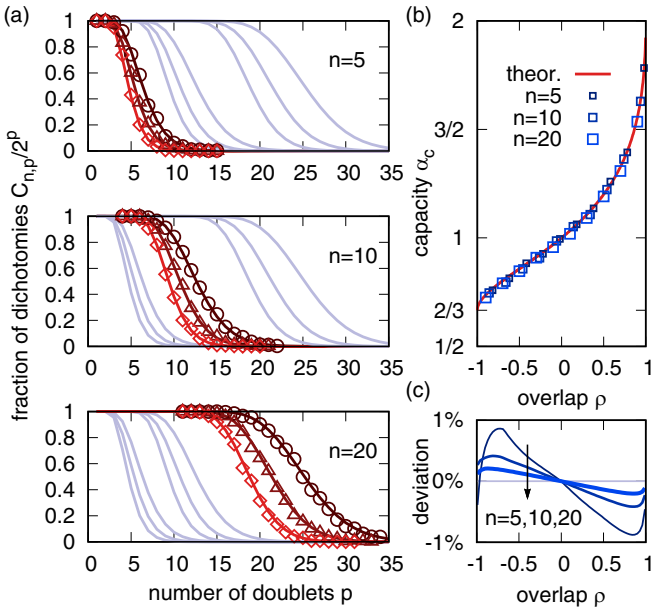

FIG. 2. Theory (solid lines) vs numerical results for $k = 2$, obtained by training a linear classifier with the perceptron algorithm. (a) Fraction of admissible dichotomies (y axis) as a function of number of doublets (x axis) in dimensions $n = 5, 10, 20$ for different values of the overlap $\rho = 0.6$ ($\circ$), 0.2 ($\triangle$), -0.2 ($\diamond$). The theoretical curves are given by Eq. (13); gray lines are just for comparing the three values of n on the same graph. (b) The capacity [Eq. (17)] (y axis) as a function of the overlap ρ (x axis). (c) Finite-size deviation of the capacity [obtained by solving numerically $C_{n,\alpha_c n} = 2^{\alpha_c n-1}$ at fixed n with $C_{n,p}$ given by Eq. (13)] from the large-n prediction Eq. (17). [Each point in (a) is a fraction over 1000 independent trials; the capacity in (b) is obtained by linearly interpolating data such as those in (a).]

An estimate for the capacity, valid for large n, can be obtained by approximating Eq. (13) as

$$C_{n,p} \approx 2 \sum_{i=0}^{n-1} K_{i,p}. \tag{14}$$

The capacity α_c is such that

$$C_{p/\alpha_c,p} \approx 2^{p-1}, \tag{15}$$

i.e., it corresponds to the value of n for which the sum of $K_{i,p}$ takes half its maximum value. The quantity $K_{i,p}$ can be interpreted as the partition function of an ensemble of directed random walks $\{\gamma_j\}_{j=1,\dots,p}$ of $p-1$ steps, with the same boundary conditions as for $k = 1$, and the following transition probabilities: $P(\gamma_j \to \gamma_j) = \Psi_2/2$, $P(\gamma_j \to \gamma_j + 1) = 1/2$, $P(\gamma_j \to \gamma_j + 2) = (1 - \Psi_2)/2$. The normalization factor 2 at the denominator is the sum of the weights Ψ_2, 1, and $1 - \Psi_2$. The capacity therefore corresponds to the median of the distribution function of the walk's end point i. We approximate the median with the mean

$$\bar{\imath} = (p-1) \sum_{l=0}^{2} l P(\gamma_j \to \gamma_j + l), \tag{16}$$

which evaluates to $\bar{\imath} = (3/2 - \Psi_2)(p-1)$, and finally we obtain

$$\alpha_c \approx \frac{p-1}{\bar{\imath}} = \frac{2}{3 - 2\Psi_2(\rho)}. \tag{17}$$

This result, with Ψ_2 given by Eq. (6), was found in Ref. [40] by means of replica calculations, and appeared more recently in other contexts in Refs. [26,32]. Our derivation is somewhat more elementary, and naturally highlights the role of the geometric quantity $\Psi_2(\rho)$.

Figure 2 compares the analytical formulas (13) and (17) with numerical results obtained by training a linear classifier with random doublets at varying dimension n, number of points p, and overlap ρ. We employ the standard Perceptron algorithm, stopping whenever a solution is found or a fixed maximum number of iterations (here 10^4) is reached; this introduces a small systematic underestimation of $C_{n,p}$, which we checked was smaller than the statistical fluctuations. Equation (13) matches perfectly as expected. Equation (17) is surprisingly precise even at very small sizes; deviations are less than 1% already for $n = 5$.

V. POLYTOPES (MULTIPLETS, GENERIC k)

Let us now move to the general case where the data are structured in multiplets of k points. We consider dichotomies of k sets of points $\xi^\mu = \{\xi_1^\mu, \dots, \xi_p^\mu\}$, with $\mu = 1, \dots, k$. The ith multiplet is the set $\xi_i = \{\xi_i^1, \dots, \xi_i^k\}$. A dichotomy ϕ is admissible if the images of all k partner points in each multiplet are equal: $\phi(\xi_i^\mu) = \phi(\xi_i^\nu)$ for all $\mu, \nu = 1, \dots, k$, separately for all $i = 1, \dots, p$. For clarity, we denote the number of admissible dichotomies by $C_{n,p}^{(k)}$, as shown in Fig. 1.

A recursion relation for $C_{n,p}^{(k)}$ can be obtained by carefully extending the method used for the doublet case. At the $(p+1)$th step, we consider the multiplet ξ_{p+1}, composed of the k points $\xi_{p+1}^1, \dots, \xi_{p+1}^k$. Let us exclude momentarily the point ξ_{p+1}^1, and suppose we know how to apply Cover's method to the set of $k-1$ points:

$$\bar{\xi}_{p+1} = \left\{\xi_{p+1}^2, \dots, \xi_{p+1}^k\right\} \subset \xi_{p+1}. \tag{18}$$

This would give an expression; let us call it

$$Q^{k-1}\left(C_{n,p}^{(k)}, C_{n-1,p}^{(k)}, \dots, C_{n-k+1,p}^{(k)}\right). \tag{19}$$

The fact that Q^{k-1} is a function of $C_{n-l,p}^{(k)}$ with $l = 0, \dots, k-1$ will be clear in the following. Intuitively, the case $k = 1$ involves only $l = 0$ and 1, the case $k = 2$ adds $l = 2$ because it uses the expression for $k = 1$ in $n-1$ dimensions, and the same pattern repeats inductively up to $k-1$ points.

The quantity Q^{k-1} represents the number of dichotomies of the set $\xi_1 \cup \xi_2 \cup \cdots \cup \xi_p \cup \bar{\xi}_{p+1}$ that are admissible on the first p multiplets [meaning that $\phi(\xi_i^\mu) = \phi(\xi_i^\nu)$ for all $\mu, \nu = 1, \dots, k$ and all $i = 1, \dots, p$] and admissible on the $k-1$ points in $\bar{\xi}_{p+1}$ [meaning that $\phi(\xi_{p+1}^\mu) = \phi(\xi_{p+1}^\nu)$ for all $\mu, \nu = 2, \dots, k$]. A number $R_{n,p}^{k-1}$ of these dichotomies are realizable by a hyperplane passing through the excluded point ξ_{p+1}^1, and are therefore all admissible. Of the remaining $Q^{k-1}(\dots) - R_{n,p}^{k-1}$ ones, a fraction $\tilde{\Psi}_k$ assign the same value to ξ_{p+1}^1 and to the points in $\bar{\xi}_{p+1}$, and are therefore admissible on

the whole multiplet ξ_{p+1}. Therefore,

$$C^{(k)}_{n,p+1} = \tilde{\Psi}_k\left[Q^{k-1}(\ldots) - R^{k-1}_{n,p}\right] + R^{k-1}_{n,p}. \tag{20}$$

While Ψ_2 was a probability (over all possible hyperplanes), $\tilde{\Psi}_k$ is a conditional probability, namely, the probability that a uniform vector w on the sphere S^{n-1} does not separate the multiplet ξ_{p+1}, conditioned on the event that w does not separate the set $\bar{\xi}_{p+1}$:

$$\tilde{\Psi}_k = \frac{\int_{S^{n-1}} dw \prod_{\mu,\nu=1}^{k} \theta\left(w\cdot\xi^{\mu}_{p+1} w\cdot\xi^{\nu}_{p+1}\right)}{\int_{S^{n-1}} dw \prod_{\mu,\nu=2}^{k} \theta\left(w\cdot\xi^{\mu}_{p+1} w\cdot\xi^{\nu}_{p+1}\right)}. \tag{21}$$

The dependence of $\tilde{\Psi}_k$ on the relative positions of the points is discussed in the Appendix, where it is shown that (i) the calculation of $\tilde{\Psi}_k$ can be reduced from n-dimensional to k-dimensional integrals and (ii) $\tilde{\Psi}_k$ depends on n only through the $k(k-1)/2$ overlaps $\rho_{\mu\nu}$ between the points in a multiplet, which we fix for all multiplets:

$$\rho_{\mu\nu} = \xi^{\mu}_i \cdot \xi^{\nu}_i, \quad i = 1,\ldots,p; \quad \mu,\nu = 1,\ldots,k. \tag{22}$$

This property allows us to treat $\tilde{\Psi}_k$ as a constant in the recursions, thus simplifying the computations. Note that, since it is a conditional probability, $\tilde{\Psi}$ can be written as a ratio of probabilities:

$$\tilde{\Psi}_k(\{\rho_{\mu\nu}\}_{\mu,\nu=1,\ldots,k}) = \frac{\Psi_k(\{\rho_{\mu\nu}\}_{\mu,\nu=1,\ldots,k})}{\Psi_{k-1}(\{\rho_{\mu\nu}\}_{\mu,\nu=2,\ldots,k})}, \tag{23}$$

where Ψ_k depends on $k(k-1)/2$ overlaps between k points, and denotes the fraction of hyperplanes not separating the k points. This definition, together with the identity $\Psi_1 = 1$, implies that the geometric quantity computed above for $k = 2$ is $\Psi_2(\rho) = \tilde{\Psi}_2(\rho)$.

The number $R^{k-1}_{n,p}$ can be obtained by applying again Cover's method with respect to the set $\bar{\xi}_{p+1}$ this time in $n-1$ dimensions because the hyperplane is constrained to pass through ξ^1_{p+1}. Hence,

$$R^{k-1}_{n,p} = Q^{k-1}\left(C^{(k)}_{n-1,p}, C^{(k)}_{n-2,p}, \ldots, C^{(k)}_{n-k,p}\right). \tag{24}$$

Finally, from Eqs. (20) and (24), the recursion for $C^{(k)}_{n,p}$ is

$$C^{(k)}_{n,p+1} = Q^{k}\left(C^{(k)}_{n,p}, C^{(k)}_{n-1,p}, \ldots, C^{(k)}_{n-k,p}\right), \tag{25}$$

where the functions Q^k (having $k+1$ arguments) satisfy the recursive functional relation

$$\begin{aligned} Q^k(x_n,\ldots,x_{n-k}) = {} & \tilde{\Psi}_k Q^{k-1}(x_n,\ldots,x_{n-k+1}) \\ & + (1-\tilde{\Psi}_k) Q^{k-1}(x_{n-1},\ldots,x_{n-k}), \end{aligned} \tag{26}$$

with the boundary $Q^1(x_n, x_{n-1}) = x_n + x_{n-1}$ given by the form of Eq. (3) for a single point.

The recursion in k can be solved, thus yielding again a recursion for $C^{(k)}_{n,p+1}$ in n and p only. Let us call $\theta_k(l)$ the coefficients in the solved recursion:

$$C^{(k)}_{n,p+1} = \sum_{l=0}^{k} \theta_k(l) C^{(k)}_{n-l,p}. \tag{27}$$

Equation (26) then becomes

$$\theta_k(l) = \tilde{\Psi}_k \theta_{k-1}(l) + (1-\tilde{\Psi}_k)\theta_{k-1}(l-1), \tag{28}$$

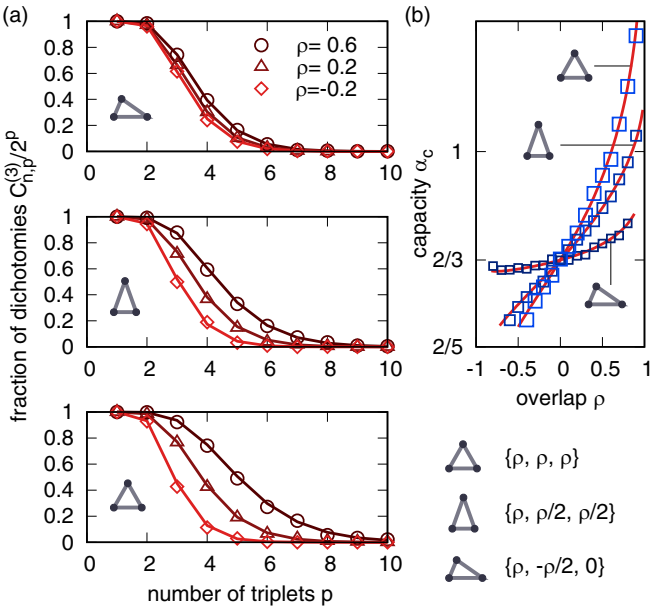

FIG. 3. Theory (solid lines) vs numerical results (symbols) for $k = 3$. (a) Fraction of admissible dichotomies as a function of the number of triplets p for triplets with overlaps $\{\rho, \rho, \rho\}$ (bottom), $\{\rho, \rho/2, \rho/2\}$ (middle), $\{\rho, -\rho/2, 0\}$ (top). Circles, triangles, and rotated squares correspond to three different values of ρ. The theoretical curves are obtained by solving numerically the recursion, Eq. (29). (b) Capacity as a function of ρ [Eq. (33)] for the same three geometries (the range of ρ is restricted by the spherical constraint).

with boundaries $\theta_1(0) = \theta_1(1) = 1$ and $\theta_k(-1) = \theta_k(k+1) = 0$. For instance, setting $k = 2$ in Eqs. (27) and (28) recovers the recursion for doublets, Eq. (9), as expected. For $k = 3$ one obtains

$$\begin{aligned} C^{(3)}_{n,p+1} = {} & \tilde{\Psi}_3\Psi_2 C^{(3)}_{n,p} + [\tilde{\Psi}_3 + \Psi_2(1-\tilde{\Psi}_3)] C^{(3)}_{n-1,p} \\ & + [\tilde{\Psi}_3(1-\Psi_2) + (1-\tilde{\Psi}_3)] C^{(3)}_{n-2,p} \\ & + (1-\tilde{\Psi}_3)(1-\Psi_2) C^{(3)}_{n-3,p}. \end{aligned} \tag{29}$$

In the process of deriving the foregoing recursion relations we considered the points ξ^{μ}_{p+1} in a particular order, therefore explicitly breaking invariance under permutations within the multiplets. We restore the invariance *a posteriori*, by prescribing that all $\tilde{\Psi}_l$ (with $l \leqslant k$) be symmetrized with respect to all $k(k-1)/2$ overlaps. For instance, when $k = 3$, the $\Psi_2 = \tilde{\Psi}_2$ appearing in Eq. (29) is to be intended as $[\Psi_2(\rho_{12}) + \Psi_2(\rho_{13}) + \Psi_2(\rho_{23})]/3$. The goodness of this prescription is substantiated by the numerical results shown in Fig. 3; see also the limit case (ii) below.

The solution for $C_{n,p}$ (with the appropriate boundary conditions) can be obtained, for instance, via generating functions, but we do not give it here. Instead, we focus on the capacity, which can be computed by the same approximate method used for $k = 2$ [Eqs. (16) and (17)]:

$$\alpha_c = \frac{\sum_{l=0}^{k}\theta_k(l)}{\sum_{l=0}^{k} l\,\theta_k(l)} = \frac{\lambda_0(k)}{\lambda_1(k)}, \tag{30}$$

where we have defined the moments

$$\lambda_m(k) = \sum_{l=0}^{k} l^m \theta_k(l). \tag{31}$$

Summing Eq. (28) over l shows that $\lambda_0(k) = \lambda_0(k-1)$ and therefore $\lambda_0(k) = \lambda_0(1) = 2$. By multiplying Eq. (28) by l and summing over l, one obtains $\lambda_1(k) = \lambda_1(k-1) + (1 - \tilde{\Psi}_k)\lambda_0(k-1)$. The boundary condition $\lambda_1(1) = 1$ then fixes the solution

$$\lambda_1(k) = 2k - 1 - 2\sum_{l=2}^{k} \tilde{\Psi}_l. \tag{32}$$

Finally, substituting $\lambda_0(k)$ and $\lambda_1(k)$ into Eq. (30) yields a remarkably simple formula for the capacity:

$$\alpha_c = \left(k - \frac{1}{2} - \sum_{l=2}^{k} \tilde{\Psi}_l \right)^{-1}. \tag{33}$$

Figure 3 compares our theory with numerical computations in the case of triplets ($k = 3$), for triangles with three, two, and no sides of the same length. The agreement is excellent. The function $\tilde{\Psi}_3$ is a double integral (given in the Appendix), which we evaluate numerically.

We mention three simple limit cases of Eq. (33).

(i) If all the points in each multiplet coincide, then $\tilde{\Psi}_l = 1$ for all $l = 2, \ldots, k$ and we recover the single point classic result $\alpha_c = 2$.

(ii) When $k = 3$ and two points of a triplet coincide the overlaps are $\{\rho, \rho, 1\}$. Symmetrizing $\tilde{\Psi}_3(\rho, \rho, 1)$ gives $\Psi_3(\rho, \rho, 1)[2/\Psi_2(\rho) + 1/\Psi_2(1)]/3$ where $\Psi_3(\rho, \rho, 1)$ is the fraction of hyperplanes not separating the three points. Clearly $\Psi_3(\rho, \rho, 1) = \Psi_2(\rho)$, and one recovers Eq. (17) for $k = 2$ as expected.

(iii) If $\tilde{\Psi}_l = 0$ for all $l = 2, \ldots, k$ Eq. (33) gives $\alpha_c = 2/(2k-1)$. This prediction matches that obtained in Ref. [15] for $(k-1)$-dimensional linear manifolds. However, this turns out to be an unphysical limit in our framework, since $\tilde{\Psi}_l$ cannot be all vanishing. For instance, for $k = 3$, equilateral triplets with overlaps $\{\rho, \rho, \rho\}$ lie on a linear manifold passing through the origin when ρ takes its minimum value $\rho_\triangle = -1/2$. The same happens for isosceles triplets $\{\rho, \rho/2, \rho/2\}$ at $\rho = -\sqrt{3}$. Interestingly, the capacity evaluated at the respective minimum ρ is $\alpha_c \approx 0.46154$ for both geometries, to be compared to the value $\alpha_c = 2/5$ found for two-dimensional linear manifolds.

Another interesting, albeit less elementary, limit case would be $k \to \infty$, taken in such a way that the points generate a sphere of radius κ. Then Eq. (33) should be related to the well-known capacity with margin κ [16], which has never been obtained by combinatorial methods [11,15] (the relation to spherical perceptual manifolds was explored in Ref. [26]).

VI. DISCUSSION

The statistical mechanics of neural networks relies routinely on the simplifying assumption that inputs are chosen randomly and independently. However empirical datasets are not assembled by tossing random independent variables. For instance, data points with the same label tend to be more similar to each other. Equivalently, one could picture a dataset as a collection of extended objects, each object composed of all points having the same label. How to incorporate this geometric structure into existing theories of machine learning (both in statistical mechanics and in computer science) is an open problem. In particular, a relevant question is how to predict if a given dataset is linearly separable, i.e., if all possible dichotomies of its classes are linearly realizable. Function counting is a way to quantify the linear separability of a dataset, via the number of binary classifications that can be realized by a linear separator. This number is a measure of the expressivity of the classifier on that particular dataset. Current theories estimate this number only with worst-case estimates, or in the unrealistic case of random uncorrelated data. The theory we have developed is a first step towards the goal of predicting the typical-case linear separability of a dataset by taking into account explicitly its structure. In particular, it allows the estimation of the capacity and expressivity of a linear classifier on a specific dataset, knowing only the overlaps between equally labeled inputs (we give an example of such an application below).

Our extension of Cover's combinatorial technique to structured data allows us to obtain closed expressions of $C_{n,p}^{(k)}$ at finite n and p, for any k [we have written explicitly the result for $k = 2$ in Eq. (13)]. Beside this, our main result is Eq. (33), which expresses the capacity as a simple function of the quantities $\tilde{\Psi}_l$. Regarding these quantities, the merit of our method is twofold: first, the $\tilde{\Psi}_l$'s are revealed to be the only relevant parameters characterizing the linear separability of the multiplets; second, they have a very simple geometric interpretation in terms of probabilities. We used here a particular parametrization of the geometric structure within the multiplets, namely, the overlaps between the different input variants. However, the theory naturally gives rise to a separation between the combinatorial aspects, encoded in the recursive relations, and the description of the geometric structure, which appears only through $\tilde{\Psi}_l$. Such modularity suggests that other metrics of data structure, possibly more adapted to empirical data in particular domains, could be conveniently incorporated into the framework.

Although the theory is generic in k, we have concentrated here on small multiplets. Besides the easier tractability, this small-k regime may be the relevant one empirically; in fact, it is well known that empirical data most often lie on manifolds the intrinsic dimension of which is much lower than that of the embedding space [41,42]. This property even appears to underpin some peculiar properties of the learning process in multilayer networks [36].

It is important to point out two limitations of our computations. First, in the spirit of mean-field calculations in statistical physics, the results should not be considered exact in the mathematical sense, in contrast with Cover's original work. Still, the excellent agreement with numerical computations suggests that at least some of the results concerning $C_{n,p}^{(k)}$ may be stated in theorem form and proved rigorously, provided the appropriate conditions can be identified. One condition likely deserving more scrutiny is the general position of the inputs. Both the mean-field theory that we have developed and the simulations reported in the figures assume that $C_{n,p}$ is averaged over the statistical ensemble. Notice that we made

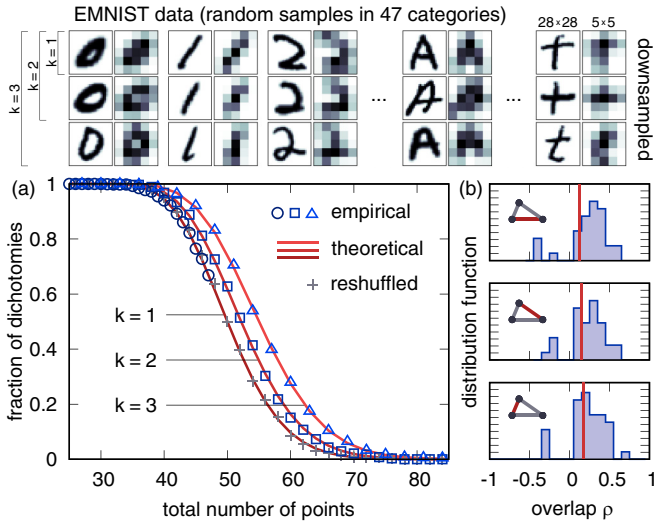

FIG. 4. Multiplets with widely different overlaps yield curves in agreement with the theoretical predictions obtained by using solely the mean overlaps. The EMNIST dataset was sampled to obtain three random images from each of its 47 categories. To grant access to the nontrivial part of the curves (where $C_{n,p}^{(k)} < 1$), images were downscaled bilinearly from 28^2 to 5^2 pixels. Low-intensity white noise was added to ensure the inputs were in general position. (a) For each value of p, we counted the fraction of times (out of 1000) that a perceptron was able to learn a random admissible dichotomy of $k = 3$ images (respectively, $k = 2, 1$) in p randomly chosen different categories. Open symbols are the numerical results; lines are the theoretical predictions for the mean overlaps [red vertical lines in (b)]. Crosses were obtained by including both admissible and nonadmissible dichotomies (at $k = 2$), equivalent to a reshuffling of the labels. Reshuffling removes the data structure, thus recovering the unstructured $k = 1$ result. Note that the x axis is the total number of points kp. (b) The empirical probability distribution functions of ρ_{12}, ρ_{23}, ρ_{13}, the overlaps between the three different images in each category.

another approximation in deriving the capacity, Eqs. (17) and (33), by substituting the mean of $P(\gamma_j \to \gamma_j + l)$ for its median [see the brief discussion above Eq. (16)]. This approximation is expected to be irrelevant in the large-n limit, as is supported by the numerical results in Fig. 2(c).

Second, the statistical ensemble we considered is somewhat restrictive, in that it assumes that the multiplets are monodisperse, meaning that they all have the same fixed set of overlaps $\{\rho_{\mu\nu}\}$. This assumption would seem to undercut the empirical applicability of the framework, considering that empirical data are always polydisperse. To address this weakness, we have performed a comparison using the Extended MNIST (EMNIST) dataset [43], an extension of the MNIST dataset containing also lowercase and uppercase letters. This dataset has the advantage of comprising 47 different categories, therefore allowing for values of p in a wider range. The results are reported in Fig. 4 (details are in the caption). Despite the broad variability of the overlaps across the categories, the mean overlaps are predictive of the full curve quantifying the fraction of admissible dichotomies that are realizable by a perceptron.

The check on EMNIST is encouraging, and suggests that polydispersity of the overlaps within each object manifold has marginal effect on the number of dichotomies. However, another assumption of our mean-field theory requires that the marginal one point probabilities be uniform on the sphere, Eq. (2). Deviations from this assumption may be relevant for more complex datasets.

Other applications and extensions of the theory appear possible. The capacity is written in Eq. (30) as a combination of the zeroth and first moments, but higher-order moments can be computed similarly and give access to other useful quantities. For instance, the second moment is related to the width of the crossover region separating the regimes where $c_{n,p} \approx 1, 0$, respectively. Furthermore, it would be interesting to express our results for general (nonlinear) separating surfaces, in the same spirit of Cover's original work, and in view of useful applications.

ACKNOWLEDGMENTS

We would like to dedicate this work to the memory of Bruno Bassetti. We thank Carlo Baldassi, Curtis G. Callan, Haim Sompolinsky, and Riccardo Zecchina for useful feedback and discussions. P.R. acknowledges funding from the European Union's Horizon 2020 programme under the Marie Skłodowska-Curie Individual Fellowship Grant Agreement No. 766442.

APPENDIX: COMPUTATION OF Ψ_k

1. Computation of $\Psi_2(\rho)$

The fraction of hyperplanes assigning the same value to two points ξ and $\bar{\xi}$ is given by

$$\Psi_2 = \frac{2}{\mathcal{N}} \int d^n x \, \delta(x^2 - 1)\theta(x \cdot \xi)\theta(x \cdot \bar{\xi}). \tag{A1}$$

The normalization factor is

$$\mathcal{N} = \int d^n x \, \delta(x^2 - 1) = \Omega_n/2, \tag{A2}$$

where Ω_n is the solid angle in n dimensions. Gram-Schmidt (GS) orthonormalization of ξ and $\bar{\xi}$ yields

$$e_1 = \xi, \quad e_2 = \frac{\bar{\xi} - \rho\,\xi}{\sqrt{1-\rho^2}}, \tag{A3}$$

where $\rho = \xi \cdot \bar{\xi}$ is the overlap between the two points. Inverting Eq. (A3) gives

$$\xi = e_1, \quad \bar{\xi} = \rho\, e_1 + \sqrt{1-\rho^2}\, e_2. \tag{A4}$$

Having orthonormalized the points allows us to safely exploit the $(n-2)$-dimensional spherical symmetry of the integral in the space orthogonal to ξ_1 and ξ_2, and to reduce it to an integral over the two-dimensional solid angle:

$$\begin{aligned}\Psi_2 &= \int \frac{d\Omega_2}{\pi} \theta(\cos\phi)\theta(\rho\cos\phi + \sqrt{1-\rho^2}\sin\phi) \\ &= \frac{2}{\pi} \arctan\left(\sqrt{\frac{1+\rho}{1-\rho}}\right),\end{aligned} \tag{A5}$$

which evaluates to the result in Eq. (6), and shows that $\Psi_2 = \Psi_2(\rho)$.

2. Computation of $\Psi_3(\rho_{12}, \rho_{13}, \rho_{23})$

Equation (23) expresses the conditional probability $\tilde{\Psi}_k$ in terms of the probabilities Ψ_k. Ψ_k is defined as the fraction of hyperplanes assigning the same value to the k points $\xi_1, \xi_2, \ldots, \xi_k$:

$$\Psi_k = \frac{2}{\mathcal{N}} \int d^n x \; \delta(x^2 - 1) \prod_{\mu=1}^{k} \theta(x \cdot \xi^\mu), \tag{A6}$$

with $\mathcal{N}$ given by Eq. (A2). For $k = 3$, the Gram-Schmidt procedure gives

$$e_1 = \xi^1, \quad e_2 = \frac{\xi^2 - \rho_{12}\xi^1}{\sqrt{1-\rho_{12}^2}}, \quad e_3 = \frac{\xi^3 - \rho_{13}e_1 - g e_2}{\sqrt{1-\rho_{13}^2 - g^2}},$$

where $\rho_{\mu\nu} = \xi^\mu \cdot \xi^\nu$ are the overlaps, and $g = (\rho_{23} - \rho_{12}\rho_{13})/\sqrt{1-\rho_{12}^2}$. Again, thanks to the spherical symmetry in the space orthogonal to the ξ^μ's the result is an integral over the three-dimensional solid angle:

$$\Psi_3 = \frac{\Gamma\left(\frac{3}{2}\right)}{\pi^{\frac{3}{2}}} \int d\Omega_3 \, \theta\left(\rho_{12}x_1 + \sqrt{1-\rho_{12}^2}x_2\right)\theta(x_1) \times \theta\left(\rho_{13}x_1 + g x_2 + \sqrt{1-\rho_{13}^2-g^2}x_3\right), \tag{A7}$$

where the measure $d\Omega_3$ can be expressed via the angles ϕ_1 and ϕ_2, and $x_1 = \sin\phi_1 \cos\phi_2$, $x_2 = \sin\phi_1 \sin\phi_2$, and $x_3 = \cos\phi_1$. As above, this computation shows that $\Psi_3 = \Psi_3(\rho_{12}, \rho_{13}, \rho_{23})$. The results presented in Fig. 3 have been obtained by integrating numerically Eq. (A7).

The procedure for $k = 2, 3$ can be extended to $k > 3$. The final result has the following structure:

$$\Psi_k(\{\rho_{\mu\nu}\}) = \frac{\Gamma\left(\frac{k}{2}\right)}{\pi^{\frac{k}{2}}} \int d\Omega_k(\phi_1, \phi_2, \ldots, \phi_k) \prod_{\alpha=1}^{k} \theta[v_\alpha(\phi)],$$

where the functions v_α appearing in the θ's can be systematically derived in a similar way from the GS procedure. This shows that $\tilde{\Psi}_k$, related to Ψ_k by Eq. (23), depends in general on the ξ^μ's only through the overlaps $\rho_{\mu\nu}$, and it can be written in terms of k-dimensional integrals.

[1] A. Krizhevsky, I. Sutskever, and G. E Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, New York, 2012), pp. 1097–1105.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Curran Associates, New York, 2014), pp. 2672–2680.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT, Cambridge, MA, 2016).

[4] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning requires rethinking generalization, Proceedings of the International Conference on Learning Representations (2017) (unpublished).

[5] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. USA **113**, E7655 (2016).

[6] M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. B. Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli, Comparing dynamics: Deep neural networks versus glassy systems, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 80, edited by J. Dy and A. Krause (PMLR, Sweden, 2018), pp. 314–323.

[7] R. Shwartz-Ziv and N. Tishby, Opening the black box of deep neural networks via information, arXiv:1703.00810.

[8] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, A closer look at memorization in deep networks, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70 (PMLR, Sydney, 2017), pp. 233–242.

[9] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, IEEE Trans. Electron. Comput. **EC-14**, 326 (1965).

[10] N. Brunel, V. Hakim, P. Isope, J.-P. Nadal, and B. Barbour, Optimal information storage and the distribution of synaptic weights: Perceptron versus Purkinje cell, Neuron **43**, 745 (2004).

[11] A. Engel and C. P. L. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, New York, 2001).

[12] J. Hertz, R. G. Palmer, and A. S. Krogh, *Introduction to the Theory of Neural Computation*, 1st ed. (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1991).

[13] M. Opper and W. Kinzel, *Statistical Mechanics of Generalization* (Springer, New York, 1996).

[14] S. Ganguli and H. Sompolinsky, Statistical Mechanics of Compressed Sensing, Phys. Rev. Lett. **104**, 188701 (2010).

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

[16] E. Gardner, Maximum storage capacity in neural networks, Europhys. Lett. **4**, 481 (1987).

[17] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A: Math. Gen. **21**, 271 (1988).

[18] C. Baldassi, E. M. Malatesta, and R. Zecchina, Properties of the Geometry of Solutions and Capacity of Multilayer Neural Networks with Rectified Linear unit Activations, Phys. Rev. Lett. **123**, 170602 (2019).

[19] B. Li and D. Saad, Exploring the Function Space of Deep-Learning Machines, Phys. Rev. Lett. **120**, 248301 (2018).

[20] J. B. Tenenbaum, V. de Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science **290**, 2319 (2000).

[21] H. Sebastian Seung and D. D. Lee, The manifold ways of perception, Science **290**, 3 (2000).

[22] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science **290**, 2323 (2000).

[23] M. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.

[24] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng, Measuring invariances in deep networks, in *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, New York, 2009), pp. 646–654.

[25] F. Anselmi, J. Z. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, and T. Poggio, Unsupervised learning of invariant representations, Theor. Comput. Sci. **633**, 112 (2016).

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

[27] SueYeon Chung, U. Cohen, H. Sompolinsky, and D. D. Lee, Learning data manifolds with a cutting plane method, Neural Comput. **30**, 3 (2018).

[28] J. O'Keefe and J. Dostrovsky, The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat, Brain Res. **34**, 171 (1971).

[29] S. Cocco, R. Monasson, L. Posani, S. Rosay, and J. Tubiana, Statistical physics and representations in real and artificial neural networks, Physica A **504**, 45 (2018).

[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in Proceedings of the Second International Conference on Learning Representations, 2014 (unpublished).

[31] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, Sensitivity and generalization in neural networks: An empirical study, Proceedings of the International Conference on Learning Representations, 2018 (unpublished).

[32] F. Borra, M. C. Lagomarsino, P. Rotondo, and M. Gherardi, Generalization from correlated sets of patterns in the perceptron, J. Phys. A: Math. Theor. **52**, 384004 (2019).

[33] T. Y. Pang and S. Maslov, Universal distribution of component frequencies in biological and technological systems, Proc. Natl. Acad. Sci. USA **110**, 3 (2013).

[34] A. Mazzolini, M. Gherardi, M. Caselle, M. C. Lagomarsino, and M. Osella, Statistics of Shared Components in Complex Component Systems, Phys. Rev. X **8**, 021023 (2018).

[35] M. Mézard, Mean-field message-passing equations in the hopfield model and its generalizations, Phys. Rev. E **95**, 022117 (2017).

[36] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modelling the influence of data structure on learning in neural networks, arXiv:1909.11500.

[37] C. Li, H. Farkhoor, R. Liu, and J. Yosinski, Measuring the intrinsic dimension of objective landscapes, Proceedings of the International Conference on Learning Representations, 2018 (unpublished).

[38] C. Guo and F. Berkhahn, Entity embeddings of categorical variables, arXiv:1604.06737.

[39] V. N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities, Theory Probab. Appl. **16**, 264 (1971).

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

[41] V. Erba, M. Gherardi, and P. Rotondo, Intrinsic dimension estimation for locally undersampled data, Sci. Rep. **9**, 17133 (2019).

[42] E. Facco, M. d'Errico, A. Rodriguez, and A. Laio, Estimating the intrinsic dimension of datasets by a minimal neighborhood information, Sci. Rep. **7**, 12140 (2017).

[43] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, EMNIST: An extension of MNIST to handwritten letters, arXiv:1702.05373.