Beyond the Storage Capacity: Data-Driven Satisfiability Transition

Pietro Rotondo,[1,2] Mauro Pastore,[1,2] and Marco Gherardi[2,1,*]
[1]Istituto Nazionale di Fisica Nucleare, sezione di Milano, via Celoria 16, 20133 Milano, Italy
[2]Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy

(Received 24 May 2020; revised 22 July 2020; accepted 13 August 2020; published 14 September 2020)

Data structure has a dramatic impact on the properties of neural networks, yet its significance in the established theoretical frameworks is poorly understood. Here we compute the Vapnik-Chervonenkis entropy of a kernel machine operating on data grouped into equally labeled subsets. At variance with the unstructured scenario, entropy is nonmonotonic in the size of the training set, and displays an additional critical point besides the storage capacity. Remarkably, the same behavior occurs in margin classifiers even with randomly labeled data, as is elucidated by identifying the synaptic volume encoding the transition. These findings reveal aspects of expressivity lying beyond the condensed description provided by the storage capacity, and they indicate the path towards more realistic bounds for the generalization error of neural networks.

DOI: 10.1103/PhysRevLett.125.120601

Introduction.—The success of deep learning has transformed data science profoundly in the last decade, within and outside physics [1–3]. In spite of the accomplishments in practical applications, we are currently facing a lack of fundamental theoretical understanding in the field [4,5]. Outstanding open questions concern the surprising effectiveness of stochastic gradient descent, which is capable of finding good minima in complex energy landscapes, and the identification of informative metrics to predict the performances of deep (many small layers) and shallow (few large layers) neural networks [6–8]. Particularly troublesome is the apparent incompatibility, within the accepted mathematical theories, between the expressive power and the generalization abilities of neural networks: ultimately, the reason why deep architectures with millions of parameters generalize well is mostly unknown [9–13].

A natural frame for these issues is statistical learning theory [14], which provides upper bounds to the probability of observing a large generalization error from a learning model with a given complexity. These bounds are often distribution independent, i.e, they are uniform in the generative model for the training data. The downside of their universality is their tendency to be too loose to be useful in practice. New measures of complexity are being studied to fill this gap, and the urgency of formulating data-dependent theories is widely expressed in the computer science literature [15–18].

While mathematical bounds usually address worst-case generalization, the main originality of the statistical physics approach is the analysis of the typical case; the distribution of the training data is therefore always an explicit ingredient of the computations. However, since the classic work of Gardner [19], data distribution has been regularly assumed to be factorized between the inputs and their labels (with the important exception of the so-called teacher-student scenario [20–22] where, nonetheless, the inputs are usually independent identically distributed random variables), thus leaving no room for their dependence, which is in essence what we call "data structure" here. This attitude is changing, and there is now a surge of interest towards the role of data in machine learning, with the goal of quantifying the extent to which the specificities of a dataset affect the performance of data-science methods and learning algorithms [23–31].

The main objective of this Letter is to investigate the effect that data structure has on the model complexity of simple architectures in machine learning. Previous research in the physics literature addressed this question via the traditional concept of storage capacity α_c, which measures the maximum load α (number of data points over number of parameters) that a model can learn with probability 1 in the thermodynamic limit. By viewing supervised learning as a constraint satisfaction problem, capacity corresponds to the transition between a satisfiable (SAT) and an unsatisfiable (UNSAT) phase, above which perfect training accuracy is achievable with probability 0. (It is worth remarking that this transition disappears if labels are provided by another, less expressive, neural network, as in the teacher-student framework.) Here we show that the compact description of learning provided by the capacity hides important detail about the model, related to its expressive power on structured data. Our point of originality is the shift from the capacity to a quantity borrowed from the foundations of statistical learning theory: the Vapnik-Chervonenkis (VC) entropy. We show that the VC entropy is nonmonotonic as a function of the load, and decreases asymptotically, at variance with the data-agnostic setting. This also contrasts with the classic bounds in statistical learning theory, which

are mostly obtained by upper bounding the VC entropy with quantities that grow polynomially in the size of the training set [32,33]. The hallmark of this nonmonotonic behavior is an additional phase transition above the storage capacity. The new critical point signals the entrance into the UNSAT phase of another satisfiability problem, related to data structure.

Cover's computation.—The VC entropy measures the expressive power of a classifier via the number of distinct dichotomies of the input data that the model can represent. A dichotomy is a function taking values in $\{0, 1\}$; equivalently, it is a classification of the input data in two groups. In principle, the VC entropy could give rise to informative bounds on the generalization error (the average number of errors on the test set), but it is usually very difficult to compute explicitly, thus statistical learning theory resorts to more accessible complexity measures.

Kernel architectures are a notable exception. Here the original input $x \in \mathbb{R}^d$ is mapped to a larger n-dimensional feature vector via a fixed nonlinear function $\phi: \mathbb{R}^d \to \mathbb{R}^n$ (for instance, quadratic kernels map the input in a $d(d+1)/2$-dimensional space via a function $\phi^{(2)}$ with components $\phi^{(2)}_{ij}(x) = x_i x_j$, with $i, j = 1, \ldots, d$ and $j \le i$).

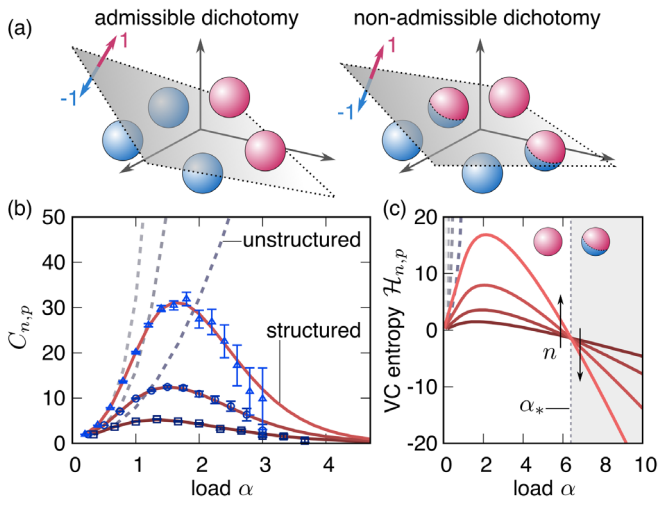

FIG. 1. (a) Input data are structured as groups of points sharing the same label (pink $= +1$, blue $= -1$). Each sphere denotes, in a stylized way, a group of points. Singly colored spheres contribute to admissible dichotomies; conversely, a dichotomy containing a doubly colored sphere is not admissible. (b),(c) The VC entropy $\mathcal{H}_{n,p}$ is the logarithm of the number $C_{n,p}$ of expressible dichotomies such that no two points belonging to the same group are classified differently. $C_{n,p}$ and $\mathcal{H}_{n,p}$ are monotonic in the load for unstructured data and nonmonotonic for structured data. Solid lines are the theory for pairs of points ($k = 2$); dashed lines are Cover's result ($k = 1$); from bottom to top, $n = 3, 4, 5$ in (b) and $n = 5, 10, 20, 40$ in (c); symbols are numerical estimates. The VC entropies at different values of n intersect roughly at the same load α_*, which separates two phases, where admissible dichotomies are asymptotically present or absent.

The VC entropy of kernel machines was obtained analytically in a remarkable paper by Cover more than half a century ago [34]. Cover calculated the number $C_{n,p}$ of dichotomies as a function of the number p of data points and the dimension n; the VC entropy is $\mathcal{H}_{n,p} = \log C_{n,p}$. In the thermodynamic limit, i.e., $n, p \to \infty$ with fixed load $\alpha = p/n$, the fraction of dichotomies $C_{n,p}/2^p$ is discontinuous at the storage capacity α_c ($\alpha_c = 2$ for the spherical perceptron). Remarkably, Cover's formula holds on very mild assumptions on the actual data points; this suggests that statistical dependence between the inputs and their labels must be conceded if one is to attain data-aware estimates. Very recently the combinatorial technique devised by Cover was extended to include this type of data structure [35], allowing the computation of the number of "admissible" dichotomies, i.e., those that are compatible with the data structure [see Fig. 1(a)].

VC entropy in a simple model of data structure.—How to formulate a significant notion of data structure is a debated issue, and different descriptions are useful in different contexts [23,25,36,37]. Here we use the definition of Ref. [35]. Data points are grouped into p subsets of k points each, where the labels are the same within each subset, and the geometric relations between points in a subset are fixed. More precisely, the input set is $\Xi = \cup_{\mu=1}^{p} \Xi_\mu$, where each $\Xi_\mu = \{\xi^\mu_a\}_{a=1,\ldots,k}$ is a set ("multiplet") of k points on the unit sphere $\xi^\mu_a \in S^{n-1} \subset \mathbb{R}^n$ such that their $k(k-1)/2$ overlaps are fixed: $\xi^\mu_a \cdot \xi^\mu_b = \rho_{ab}$ for all $\mu = 1, \ldots, p$. The ensemble we consider is the flat probability measure on the kp points ξ^μ_a, conditioned to these constraints. The admissible dichotomies ϕ of Ξ are those for which $\phi(\xi^\mu_a) = \phi(\xi^\mu_b)$ for all $a, b = 1, \ldots, k$ and $\mu = 1, \ldots, p$. The usual unstructured ensemble is recovered either when $k = 1$ (where no overlaps need to be specified), or, for any k, when $\rho_{ab} = 1$ for all a, b. We stress that structure, in this definition, is not a property of the inputs or of the labels alone: it describes the relations between inputs and labels. This model of data structure is closely related to the concept of "perceptual manifolds" inspired by neuroscience [25,38], and was recognized in Ref. [31] as a promising theoretical tool to address the problem of generalization.

The average number of admissible dichotomies $C_{n,p}$ of p sets of k points (the logarithm of which is the VC entropy $\mathcal{H}_{n,p}$) satisfies the mean-field recurrence relation [35]

$$C_{n,p+1} = \sum_{l=0}^{k} \theta^k_l C_{n-l,p}. \tag{1}$$

The boundary conditions depend mildly on the geometry, but they can be approximated by $C_{n\ge 1,1} = 2$, $C_{0,p} = 0$. Each coefficient θ^k_l in Eq. (1) depends on $k-1$ numbers $\{\psi_m\}_{m=2,\ldots,k}$, with $0 \le \psi_m \le 1$, having the following geometric-probabilistic interpretation. Let $w \in S^{n-1}$ be a

random vector with the uniform measure on the unit sphere. Consider any multiplet Ξ_μ, and a subset $\Xi' \subseteq \Xi_\mu$ of $m \leq k$ points. Then ψ_m is the symmetrized probability that the scalar product $w \cdot \xi$ has the same sign for all $\xi \in \Xi'$, conditioned on it having the same sign for all $\xi \in \Xi' \setminus \{\xi_\star\}$: $\psi_m = 2\langle \Pr[(w\cdot\xi_\star) > 0 | (w\cdot\xi) > 0 \, \forall \xi \in \Xi' \setminus \{\xi_\star\}] \rangle_{\rm sym}$, where the symmetrization $\langle \cdots \rangle_{\rm sym}$ is performed by averaging over all subsets Ξ' and over all choices of $\xi_\star \in \Xi'$. These quantities can be expressed in terms of the overlaps ρ_{ab}, e.g., $\psi_2(\rho) = 2\pi^{-1} \arctan \sqrt{(1+\rho)/(1-\rho)}$.

Remarkable differences between structured and unstructured data appear if one compares numerical solutions of Eq. (1) for $k = 1$ (unstructured) and $k = 2$ (structured) (Fig. 1). The VC entropy $\mathcal{H}_{n,\alpha n}$, as a function of α at fixed n, diverges with α in the unstructured case (it does logarithmically, thus the fraction of realizable dichotomies $C_{n,\alpha n}/2^{\alpha n}$ converges to 0 for $\alpha \to \infty$). On the contrary, $\mathcal{H}_{n,\alpha n}$ is nonmonotonic in the load for structured data, and $C_{n,\alpha n}$ is itself asymptotically 0. Strikingly, curves corresponding to different values of n cross each other roughly at the same load α_*, similarly to what $C_{n,\alpha n}/2^{\alpha n}$ does around the storage capacity α_c. Hence, in the thermodynamic limit the VC entropy diverges to $+\infty$ for fixed $\alpha < \alpha_*$ and to $-\infty$ for $\alpha > \alpha_*$. As will be elucidated by the following computations, this transition is driven by a trade-off between an entropic term, related to the combinatorial growth of the number of dichotomies with the load, and an energetic term, due to the constraints that define data structure.

Transition point via combinatorial analysis.—The transition point in the thermodynamic limit is accessible by a perturbative analysis. In some cases it is possible to solve Eq. (1) explicitly, but we construct here an indirect method, based on analytic combinatorics [39]. This method has the crucial advantage of being applicable despite the fact that (i) $C_{n,p}$ is not known in closed form for generic k, and (ii) the recurrence equation itself has implicitly defined coefficients (see Ref. [40] for details on the computations, and for applications to simpler cases where the above restrictions do not apply).

Let $g_n(z)$ be the ordinary generating function of $C_{n,p}$ with respect to the variable p: $g_n(z) = \sum_{p=1}^{\infty} C_{n,p} z^p$. At fixed n, $g_n(z)$ encodes the large-p asymptotics of $C_{n,p}$ via its singular behavior. In particular, if $g_n(z)$ is a rational function the dominant pole of which is of order r and lies at $z = z_0$, with finite part $R \equiv \lim_{z\to z_0}(z_0 - z)^r g_n(z)$, then, for large p, $C_{n,p} \sim R z_0^{-p-r} B(p + r - 1, r - 1)$, where $B(a, b)$ is the binomial coefficient $\binom{a}{b}$.

Multiplying Eq. (1) by z^p and summing over p (taking care of the boundary conditions) gives a recurrence relation for $g_n(z)$:

$$g_n(z) = \frac{z}{1 - z\theta_0^k}\left[2 + \sum_{l=1}^{k} \theta_l^k g_{n-l}(z)\right], \tag{2}$$

with $g_{n\leq 0}(z) = 0$. Iteration of Eq. (2) n times, starting from the nonsingular initial condition at $n = 0$, yields a singular $g_n(z)$, whose pole, generated by the pole in the right-hand side of the recurrence relation, lies at $z_0 = 1/\theta_0^k$, has order $r = n$, and finite part $R = 2(\theta_1^k)^{n-1}(\theta_0^k)^{-2n}$. Finally, the asymptotic form of the VC entropy is $\mathcal{H}_{n,\alpha n} \sim \log C(\alpha; n)$, with

$$C(\alpha; n) = 2\frac{\Gamma(\alpha n + n)}{\Gamma(n)\Gamma(\alpha n + 1)}(\theta_1^k)^{n-1}(\theta_0^k)^{(\alpha-1)n}. \tag{3}$$

Conveniently, $C(\alpha; n)$ depends only on the first two θ_l^k's (see Ref. [40] for their expressions as functions of the probabilities ψ_m). The transition is at the point $\alpha = \alpha_*$ where the VC entropy is asymptotically constant in n, i.e., $\partial_n \mathcal{H}_{n,\alpha_* n} \to 0$. From Eq. (3) one obtains

$$S(\alpha_*) + (\alpha_* - 1)\log\theta_0^k + \log\theta_1^k = 0, \tag{4}$$

with $S(\alpha) \equiv (\alpha + 1)\log(\alpha + 1) - \alpha \log \alpha$. Equation (4) expresses the trade-off between a positive entropic term $S(\alpha)$, the same as for unstructured data, and a structure-dependent energetic term. It has two solutions: α_* is the larger.

Consider the case $k = 2$, where input data are pairs of points with fixed pairwise overlap ρ. Then $\theta_0^k = \psi_2(\rho)$, $\theta_1^k = 1$, and α_* is an increasing function of ρ. Coherently, α_* diverges when $\rho \to 1$, thus recovering the unstructured case $k = 1$, where no transition is present. Figure 2 shows that (i) the value of α_* satisfying Eq. (4) matches that obtained by numerical integration of the recursion Eq. (1), and (ii) the transition can be probed by sampling small

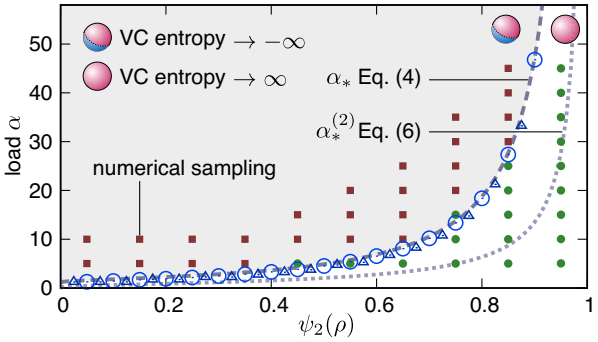

FIG. 2. Phase diagram of the VC entropy for $k = 2$. The dashed line is the theoretical prediction for α_* obtained by combinatorial methods; the dotted line is the transition line of the synaptic volume [Eq. (5)] in the annealed approximation. Empty symbols are numerical results, obtained by finding the intersection between two curves $C_{n_1,\alpha n_1}$ and $C_{n_2,\alpha n_2}$ with $n_1 = 40$, $n_2 = 20$ (circles) and $n_1 = 6$, $n_2 = 3$ (triangles); each filled symbol is obtained by sampling 10^5 random inputs with $n = 3$ (red squares = no admissible dichotomy, green circles = at least one admissible dichotomy).

random linear classifiers (see the caption). The phenomenology is the same for all k [40].

Similarly to the critical behavior at the SAT-UNSAT transition of random constraint satisfaction problems [41,42], the number of dichotomies, as a function of the reduced control parameter $\hat{\alpha} = (\alpha - \alpha_*)/\alpha_*$, obeys a finite-size scaling form $C(\alpha; n) = n^{-\beta/\nu} F(\hat{\alpha} n^{1/\nu})$, with critical exponents $\beta = 1/2$ and $\nu = 1$, where F is a regular function (see Ref. [40] for the explicit formula). At $\alpha = \alpha_*$, $C(\alpha; n)$ vanishes as a power law in the dimensionality n; the exponent ν controls the scaling of the width of the critical region (by contrast, $\nu = 2$ at the storage capacity).

Identification of the relevant synaptic volume.—The phase transition at α_* can be interpreted as the SAT-UNSAT transition of the following constraint satisfaction problem: given a realization of the "disorder" Ξ, find a vector W identifying a linearly realizable dichotomy of Ξ that is admissible. This characterization indicates that the following synaptic volume should pinpoint the transition:

$$V(\Xi) = \int D^p\sigma \int D^n W \prod_{\mu,a=1}^{p,k} \theta\left[\sigma^\mu \sum_{i=1}^{n} W_i (\xi_a^\mu)_i\right], \quad (5)$$

where $\theta[\cdots]$ is the Heaviside theta, $(\xi_a^\mu)_i$ denotes the ith component of the ath element of the μth multiplet, $D^n W$ is a shorthand for a Gaussian or spherical measure over the weights, and $D^p\sigma = \prod_\mu [\delta(\sigma^\mu - 1) + \delta(\sigma^\mu + 1)] d\sigma^\mu$. Besides the data structure, encoded in the multiplets Ξ_μ, the synaptic volume [Eq. (5)] differs from the ordinary Gardner volume by the integration over the labels σ. Intuitively, as long as $V(\Xi)$ grows exponentially with n at fixed load α, at least one classification compatible with the input-label constraints can be expressed by the model. Thus, the scaling of $V(\Xi)$ is a proxy of the nonmonotonic behavior of the VC entropy for a given data structure.

We restrict the analysis to data structured as pairs of points ($k = 2$), and we compute $V(\Xi)$ in the simplest approximation scheme, averaging at the annealed level over the inputs. (See Ref. [40] for the replica theory.) For $\rho = 1$ we recover the unstructured case: $\langle V(\Xi) \rangle$ diverges for any load α, in agreement with Cover's theory (a polynomial number of classifications can be realized by a kernel architecture). The situation changes for $\rho < 1$. In this regime data structure becomes relevant, and there appears a critical load $\alpha_*^{(2)}(\rho)$ for which the synaptic volume shrinks exponentially fast in n. Above this threshold, which is given by

$$\alpha_*^{(2)}(\rho) = -\frac{\log(2\pi) + 1}{2\log\left(1/2 + \pi^{-1}\arcsin\rho\right)}, \quad (6)$$

none of the classifications compatible with the data structure can be realized by the kernel architecture. The threshold computed in the annealed approximation provides a lower bound to the α_* evaluated by the combinatorial approach (see Fig. 2).

Margin-driven transition with unstructured data.—Margin classifiers are prominent in statistical learning theory, as their generalization error can be kept under control via the margin, and they lie at the core of the powerful idea of support vector machines [33,43]. A significant observation linking classification with margin and classification of structured data was done in Ref. [25]: linear classification with margin κ is equivalent to learning a set of spherical manifolds with radius equal to the margin. The equivalence, valid for a kernel machine with kernel φ, holds in the following sense: the set of d-dimensional weights W in feature space realizes the mapping with margin κ if and only if $\sigma^\mu = \mathrm{sign}(W \cdot \zeta^\mu)$ for all μ and all ζ^μ such that $|\zeta^\mu - \varphi(\xi^\mu)|^2 < \kappa^2$. Intuitively, the constraints of the satisfiability problem are shifted from the data to the function class. (If the margin is negative the problem is no more convex, and bears connections to jamming phenomena [44]).

This observation suggests that the VC entropy of a margin classifier with randomly labeled (i.e., unstructured) data should present the same phenomenology described above for data structured in multiplets. To our knowledge, there is no combinatorial technique to compute the entropy in this case, thus we use an integrated synaptic volume analogous to Eq. (5) as a probe into the phase transition. Again, in the annealed approximation, the volume shrinks exponentially fast above a threshold load, given by

$$\alpha_*^M(\kappa) = -\frac{\log(2\pi) + 1}{2\log \mathrm{Erfc}(\kappa)}. \quad (7)$$

As in the case of zero-margin classification of multiplets, $\alpha_*^M(\kappa) \to \infty$ when the constraints are relaxed ($\kappa \to 0$ in this case), and $\alpha_*^M(\kappa) \to 0$ when the constraints become unsatisfiable ($\kappa \to \infty$).

Discussion.—Finding compact scalar metrics descriptive of the complexity and the flexibility of a hypothesis space is a shared effort of statistical physics and statistical learning theory. Unsophisticated quantities such as the number of degrees of freedom are merely superficial indicators of the expressive power of a given model, and they fail at the task of characterizing the model's generalization properties, especially in applications to nonsynthetic datasets. This is partly true even for more refined quantities such as the VC dimension and its distribution-dependent counterparts. The importance of including data specificities in the existing frameworks is recognized in both physics and computer science. In particular, it is well appreciated that restricting the hypothesis class by imposing a margin is beneficial to generalization. A large body of work in modern statistical learning theory is devoted to proving data-dependent bounds on the generalization error.

However, these results are obtained by bounding the VC entropy with monotonically increasing functions of the sample size p. Our results suggest that, in principle, these results could be improved substantially already by including rather unrestrictive priors on the data distribution.

Here, in the spirit of statistical physics, we have focused on simple architectures and a simple implementation of data structure. This approach enabled us to obtain tractable analytical expressions that serve, in a wider context, as a proof of principle, and promote two main points: (i) The concept of storage capacity in the statistical physics of machine learning should be complemented by other, preferably data-oriented, "order parameters" of model complexity. (ii) Data structure, in the form of dependence or constraints between inputs and labels, should be investigated in the framework of statistical learning theory, acknowledging the possibility of an asymptotically decreasing VC entropy. In this Letter we reported on the discovery of a data-driven phase transition, which appears to be a good candidate for the pursuit of point (i). Point (ii) is explored in more depth in Ref. [40]. How to address these issues for deep neural networks, or even in more generality in the context of machine learning, is compelling matter for future work.

P. R. acknowledges funding from the Fellini program under the H2020-MSCA-COFUND action, Grant Agreement No. 754496, I. N. F. N. (IT).

*marco.gherardi@mi.infn.it

[1] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, Cambridge, 2016).

[3] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV,* (2016).

[4] S. Mallat, Understanding deep convolutional networks, Phil. Trans. R. Soc. A **374**, 20150203 (2016).

[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. U.S.A. **113**, E7655 (2016).

[6] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, in *On the Expressive Power of Deep Neural Networks*, Proceedings of Machine Learning Research, edited by D. Precup and Y. W. Teh (PMLR, 2017), Vol. 70, pp. 2847–2854.

[7] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, Proc. Natl. Acad. Sci. U.S.A. **115**, E7665 (2018).

[8] P. Chaudhari and S. Soatto, Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks, in 2018 Information Theory and Applications Workshop (ITA), San Diego, CA, (2018), pp. 1–10.

[9] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning requires rethinking generalization, in Proceedings of the International Conference on Learning Representations, arXiv:1611.03530.

[10] C. H. Martin and M. W. Mahoney, Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior, arXiv: 1710.09553.

[11] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, Entropy-SGD: biasing gradient descent into wide valleys, J. Stat. Mech. (2019) 124018.

[12] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, Exploring generalization in deep learning, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (Curran Associates Inc., Red Hook, NY, USA, 2017), pp. 5949–5958.

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

[14] V. N. Vapnik, The nature of statistical learning theory, *The Nature of Statistical Learning Theory* (Springer-Verlag New York, Inc., 2013).

[15] L. Bottou, Making Vapnik–Chervonenkis bounds accurate, in Measures of Complexity. Festschrift for Alexey Chervonenkis, edited by V. Vovk, H. Papadopoulos, and A. Gammerman (Springer, 2015), pp. 143–155.

[16] A. Antos, B. Kégl, T. Linder, and G. Lugosi, Data-dependent margin-based generalization bounds for classification, J. Mach. Learn. Res. **3**, 73 (2002).

[17] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, Structural risk minimization over data-dependent hierarchies, IEEE Trans. Inf. Theory **44**, 1926 (1998).

[18] D. Cohn and G. Tesauro, How tight are the Vapnik-Chervonenkis bounds?, Neural Comput. **4**, 249 (1992).

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

[20] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, 2001).

[21] E. Levin, N. Tishby, and S. A. Solla, A statistical approach to learning and generalization in layered neural networks, Proc. IEEE **78**, 1568 (1990).

[22] H. S. Seung, H. Sompolinsky, and N. Tishby, Statistical mechanics of learning from examples, Phys. Rev. A **45**, 6056 (1992).

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

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

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

[26] 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).

[27] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modelling the influence of data structure on learning in

neural networks: the hidden manifold model, arXiv:1909.11500.

[28] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, Generalisation error in learning with random features and the hidden manifold model, arXiv:2002.09339.

[29] V. Erba, S. Ariosto, M. Gherardi, and P. Rotondo, Random geometric graphs in high dimension, Phys. Rev. E **102**, 012306 (2020).

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

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

[32] O. Bousquet, S. Boucheron, and G. Lugosi, Introduction to statistical learning theory, in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, 2003, Tübingen, Germany, 2003, Revised Lectures*, edited by O. Bousquet, U. von Luxburg, and G. Rätsch (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004), pp. 169–207.

[33] V. N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Networks **10**, 988 (1999).

[34] 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).

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

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

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

[38] H. S. Seung and D. D. Lee, The manifold ways of perception, Science **290**, 2268 (2000).

[39] P. Flajolet and R. Sedgewick, *Analytic Combinatorics* (Cambridge University Press, Cambridge, 2009).

[40] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, companion paper, Statistical learning theory of structured data, Phys. Rev. E **102**, 032119 (2020).

[41] S. Kirkpatrick and B. Selman, Critical Behavior in the Satisfiability of Random Boolean Expressions, Science **264**, 1297 (1994).

[42] M. Leone, F. Ricci-Tersenghi, and R. Zecchina, Phase coexistence and finite-size scaling in random combinatorial problems, J. Phys. A **34**, 4615 (2001).

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

[44] S. Franz and G. Parisi, The simplest model of jamming, J. Phys. A **49**, 145001 (2016).