Unveiling the Structure of Wide Flat Minima in Neural Networks

Unveiling the Structure of Wide Flat Minima in Neural Networks

Carlo Baldassi,[1] Clarissa Lauditi,[2] Enrico M. Malatesta,[1] Gabriele Perugini,[1] and Riccardo Zecchina[1]
[1]Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy
[2]Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy

(Received 2 July 2021; revised 6 December 2021; accepted 8 December 2021; published 29 December 2021)

The success of deep learning has revealed the application potential of neural networks across the sciences and opened up fundamental theoretical problems. In particular, the fact that learning algorithms based on simple variants of gradient methods are able to find near-optimal minima of highly nonconvex loss functions is an unexpected feature of neural networks. Moreover, such algorithms are able to fit the data even in the presence of noise, and yet they have excellent predictive capabilities. Several empirical results have shown a reproducible correlation between the so-called flatness of the minima achieved by the algorithms and the generalization performance. At the same time, statistical physics results have shown that in nonconvex networks a multitude of narrow minima may coexist with a much smaller number of wide flat minima, which generalize well. Here, we show that wide flat minima arise as complex extensive structures, from the coalescence of minima around "high-margin" (i.e., locally robust) configurations. Despite being exponentially rare compared to zero-margin ones, high-margin minima tend to concentrate in particular regions. These minima are in turn surrounded by other solutions of smaller and smaller margin, leading to dense regions of solutions over long distances. Our analysis also provides an alternative analytical method for estimating when flat minima appear and when algorithms begin to find solutions, as the number of model parameters varies.

DOI: 10.1103/PhysRevLett.127.278301

Machine learning has undergone a tremendous acceleration thanks to the performance of deep networks [1]. Complex architectures are able to achieve unexpected performance in disparate domains, from language processing [2] to protein structure prediction [3,4], just to name a few recent impressive results. A key aspect of all these models is the nonconvex nature of the learning problem. The learning process must be able to converge in a very high-dimensional space and in the presence of a huge number of local minima of the loss function which measures the error rate on the data set. Surprisingly, this goal can be achieved by algorithms designed for convex problems with just few adjustments, such as choosing highly parametrized architectures, using dynamic regularization techniques, and choosing appropriate loss functions [5]. In practice, neural networks with hundreds of millions of variables can be successfully optimized by algorithms based on the gradient descent method [6].

The study of the geometric structure of the minima of the loss function is essential for understanding the dynamic phenomena of learning and explaining generalization capabilities. Several empirical results have shown a reproducible correlation between the so-called flatness of the minima achieved by algorithms and generalization performance [7–9]. In a sense that needs to be made rigorous, the loss functions of neural networks seem to be characterized by the existence of large flat minima that are both accessible and well generalizable [10–13]. Moreover, similar minima are found in the case of randomized labels [14] and different data sets, suggesting that they are a robust property of the networks.

This scenario is upheld by some recent studies based on statistical physics methods [15–18], which show that in tractable models of nonconvex neural networks a multitude of minima with poor generalization capabilities coexists with a smaller number of wide flat minima, always known as high local entropy minima, that generalize close to optimality [15]. These studies rely on large-deviation methods that focus on minima surrounded at a given distance by a very large number of other minima. The analytical results are corroborated by numerical studies that confirm the accessibility of wide flat minima by simple algorithms that do not try to sample from the dominating set of minima [19].

Here, we provide analytical results on the geometric structure of these wide flat minima. We take as analytically tractable nonconvex model a prototypical neural network with N binary weights performing a binary classification task, trained on $P = \alpha N$ random patterns, investigated in the thermodynamic limit of large N and large P, with $\alpha = P/N = \mathrm{O}(1)$. This model has been extensively studied with mean field statistical physics methods [20] and by rigorous techniques [21]. The solutions of the learning task (zero-error configurations) can be characterized by their robustness to local perturbations of the weights, called margin and denoted by κ. All configurations within a radius

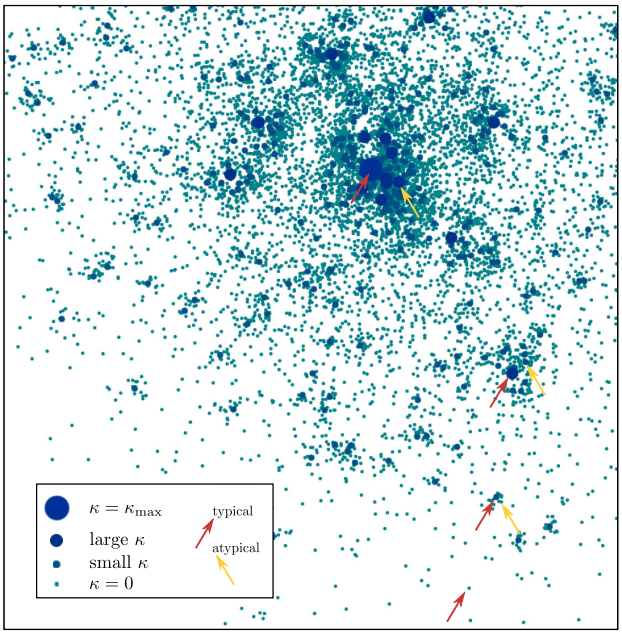

FIG. 1. Representation of a portion of the space of network configurations. The dots represent solutions (zero-error configurations) with different κ margins. Red arrows indicate four examples of typical solutions. Yellow arrows indicate three examples of the type of atypical solutions found around the typical ones with a larger margin. Low-margin solutions are more numerous than high-margin solutions. Typical low-margin solutions are isolated and distant from each other. Typical high-margin solutions are also distant from each other, but less so, and tend to be surrounded by (atypical) low-margin solutions. Thus, the higher-margin solutions are rare, but they lie within dense, extended regions that result from the coalescence of the low-margin solutions.

proportional to $\kappa\sqrt{N}$ around a κ-margin solution are also solutions. The number of κ-margin solutions is typically exponential in N, i.e., $\exp[N\phi(\alpha,\kappa)]$. Since $\phi(\alpha,\kappa)$ is monotonically decreasing with κ, high-margin solutions are exponentially rare compared to zero-margin solutions. However, they tend to concentrate in particular regions, and are in turn surrounded by other solutions of smaller and smaller margin. This coalescence of minima results in dense regions of solutions over distances of size $O(N)$. This is illustrated in Fig. 1, which shows a two-dimensional qualitative sketch of the picture that emerges from our analysis of the geometric distribution of minima, for a not-too-large value of α. As α increases, the solutions thin out, their margin gets smaller, and above some critical α the large connected structures break up and eventually disappear.

Our results provide a clearer picture regarding the internal structure of the flat minima and allow us to define an alternative analytical method for estimating the threshold at which they disappear and up to which algorithms are able find solutions efficiently. We show that, for sufficiently small values of α, the zero-error solutions have the following properties: (1) The Hamming distance between typical solutions is a rapidly decreasing function of their margin κ. Despite being exponentially less numerous (in N) compared to the $\kappa = 0$ solutions, the $\kappa > 0$ solutions tend to have small mutual distance. (2) Typical solutions with a prescribed margin $\tilde{\kappa} > 0$ are always surrounded at $O(N)$ Hamming distance by an exponential number of smaller margin solutions. By increasing $\tilde{\kappa}$, we make sure to target higher local entropy regions.

While the notion of margin has been developed in the context of shallow networks where it can be directly linked to generalization, the notion of flatness, or high local entropy, applies also to deep networks for which there is no straightforward way to define the margin for the hidden layer units. High local entropy minima are stable with respect to perturbations of the input and of the internal representations.

The model.—For simplicity, we discuss here the results of our study by considering a single-layer [22] network with N binary weights $\boldsymbol{w} \in \{-1,1\}^N$, which is perhaps the simplest to define nonconvex neural network endowed with a nontrivial geometric structure of solutions. In the Supplemental Material (SM) [23] we detail the analytical results for models with one hidden layer, which lead to a qualitatively similar geometric scenario, and also report numerical results for deep networks.

Given a (binary) pattern $\boldsymbol{\xi} \in \{-1,1\}^N$ as input to the network, the corresponding output is computed as $\sigma_{\text{out}} = \text{sign}(\boldsymbol{w}\cdot\boldsymbol{\xi})$. We consider a training set composed of $\mu = 1,\ldots,P = \alpha N$ independent identically distributed, unbiased random binary patterns $\boldsymbol{\xi}^\mu = \{-1,1\}^N$ and labels $\sigma^\mu = \{-1,1\}$ [25,26]. The learning problem consists in finding weights that realize all the input-output mappings of the training set. In this Letter, we are interested in locally robust solutions. We quantify this by imposing that for every pattern in the training set, the weights should have stability $\Delta^\mu \equiv (\sigma^\mu/\sqrt{N})\boldsymbol{w}\cdot\boldsymbol{\xi}^\mu$, larger than a given margin κ, which therefore represents the distance from the classification boundary in the direction of the correct label. The flat measure over these configurations is proportional to $\mathbb{X}_{\xi,\sigma}(\boldsymbol{w};\kappa) = \prod_{\mu=1}^{P}\Theta[(\sigma^\mu/\sqrt{N})\sum_{i=1}^{N} w_i\xi_i^\mu - \kappa]$ where $\Theta(\cdot)$ is the Heaviside theta function; this quantity is 1 if the weights $\boldsymbol{w}$ classify correctly all the patterns with margin κ, and 0 otherwise. The number of κ-margin solutions is given by

$$Z = \sum_{\{w_i=\pm 1\}} \mathbb{X}_{\xi,\sigma}(\boldsymbol{w};\kappa), \tag{1}$$

where we have dropped the dependence of Z on ξ and σ to lighten the notation. Indeed, Z is the partition function of a flat measure over the κ-margin solutions, which in turn is the zero-temperature limit of an equilibrium Gibbs measure, with the number of violated patterns as the energy. The corresponding Gibbs entropy of the solutions can be obtained as

$$\phi(\alpha,\kappa) = \lim_{N\to\infty} \frac{1}{N} \langle \ln Z \rangle_{\xi,\sigma}, \tag{2}$$

where $\langle \ldots \rangle_{\xi,\sigma}$ denotes the average over random patterns and labels. In the following, we set $\sigma^\mu = 1$ for every $\mu = 1, \ldots, P$ without loss of generality, since we can perform the transformation $\xi_i^\mu \to \sigma^\mu \xi_i^\mu$, without affecting the probability measure of the patterns. Since the model is discrete, the entropy has a lower bound of 0. In the limit of large N the model exhibits a sharp transition at the *critical capacity* $\alpha_c(\kappa)$, defined as the maximum α with nonvanishing entropy: $\phi[\alpha_c(\kappa),\kappa] = 0$. For $\alpha < \alpha_c(k)$ the probability that an instance of the problem has a solution is 1, but it sharply drops to zero beyond this threshold [27] (see also [21] for a recent rigorous proof of the value for zero margin $\alpha_c(0) \simeq 0.833$).

Distances between typical solutions.—We have computed the entropy of solutions, Eq. (2), using the replica method (details in the SM [23]). As shown in Fig. 2, we find that the Hamming distance between solutions that arises from the replica calculation is a rapidly decreasing function of the margin. The entropy is also a decreasing function of the margin (see SM [23]), meaning that solutions with a larger margin are exponentially fewer, but are much less dispersed. The closest solutions are those with maximum margin $\kappa_{\max}(\alpha)$, defined as the largest κ with nonvanishing entropy: $\phi[\alpha,\kappa_{\max}(\alpha)] = 0$.

Isolated and nonisolated solutions.—A key question is how, below the critical capacity, the solutions are arranged and how the structure of solution space affects the performance of learning algorithms. As discussed in [27,28] the structure of typical zero-margin solutions in the whole phase below $\alpha_c(\kappa = 0)$ consists of isolated clusters of vanishing entropy (so-called *frozen*-one-step replica symmetry breaking scenario). This means that one has to flip an extensive number of weights in order to find the closest solution. This scenario was recently confirmed also in simple one-hidden layer neural networks with generic activation functions [17] (but see also [29]) and also rigorously for the symmetric perceptron [30,31]. This type of landscape with point solutions would suggest that finding such solutions is a hard optimization problem, in contrast to more recent algorithmic evidence [32,33]. This seeming contradiction was resolved in [15,19,34] where it was shown that there exist rare dense regions of solutions that are accessible by simple algorithms. Subsequent work has suggested that the architectures and algorithms used for deep networks exploit the properties of these dense flat minima [16,17,35,36].

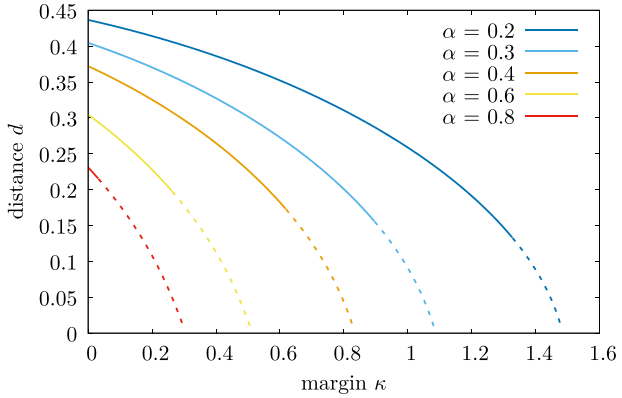

FIG. 2. Normalized Hamming distance between typical solutions as a function of their margin, for $\alpha = 0.2$, 0.3, 0.4, 0.6, and 0.8 (top to bottom). The lines become dashed when the entropy of solutions becomes negative, i.e., when $\kappa > \kappa_{\max}$ (see text).

Here, we want to understand the geometry of the dense regions, in particular how they relate to the $\kappa > 0$ solutions (see [17] for a discussion of the distribution of the stabilities inside a high-local-entropy region). We begin by analyzing in which part of the landscape high-margin solutions tend to be concentrated. Given a configuration $\tilde{\boldsymbol{w}}$, that we call the "reference," we define the *local entropy* of $\tilde{\boldsymbol{w}}$ as the logarithm, divided by N, of

$$\mathcal{N}_\xi(\tilde{\boldsymbol{w}}, d, \kappa) = \sum_{\boldsymbol{w}} \mathbb{X}_\xi(\boldsymbol{w};\kappa)\delta\left(N(1-2d) - \sum_{i=1}^{N} \tilde{w}_i w_i\right). \tag{3}$$

This expression counts the number of κ-margin solutions $\boldsymbol{w}$ which lay at normalized Hamming distance d from the reference $\tilde{\boldsymbol{w}}$. Studying the local entropy profile as we vary d allows to characterize the density of solutions (with given κ) in an extensive neighborhood of any given configuration. We are interested in describing the surroundings of typical solutions of given margin $\tilde{\kappa}$, as sampled from the Gibbs measure Eq. (1). Thanks to the self-averaging property, for sufficiently large N the local entropy profile around a typical $\tilde{w}$ can be computed as the double average over the choice of the reference solution and of the training set, i.e., by the so-called Franz-Parisi (FP) entropy [28,37]:

$$\phi_{\mathrm{FP}}(d;\alpha,\tilde{\kappa},\kappa) = \frac{1}{N}\left\langle \frac{1}{Z}\sum\nolimits_{\tilde{\boldsymbol{w}}} \mathbb{X}_\xi(\tilde{\boldsymbol{w}};\tilde{\kappa}) \ln \mathcal{N}_\xi(\tilde{\boldsymbol{w}}, d, \kappa) \right\rangle_\xi. \tag{4}$$

This quantity can be calculated using the Laplace method. We performed the calculations in the so-called replica symmetric (RS) ansatz for the order parameters, taking care to check its stability with respect to replica-symmetry-breaking effects (see SM [23] for details). Within the RS ansatz, negative entropies may appear, signaling that the number of solutions $\mathcal{N}_\xi$ is 0 [27].

We found that, for any value of α in the range $0 < \alpha < \alpha_c(\kappa)$, there are several phases depending on the value of $\tilde{\kappa}$ (shown in Fig. 3): (1) For $\tilde{\kappa} = 0$ we recover the results of [28]: $\phi_{\mathrm{FP}}(d)$ is always negative in a neighborhood of $d = 0$, meaning that the solutions are isolated, and it has only one maximum, located at the typical distance between solutions with margin $\tilde{\kappa}$ and κ). (2) When $0 < \tilde{\kappa} \le \tilde{\kappa}_{\max}(\alpha)$ there always exists a neighborhood of $d = 0$ where the average local entropy is positive,

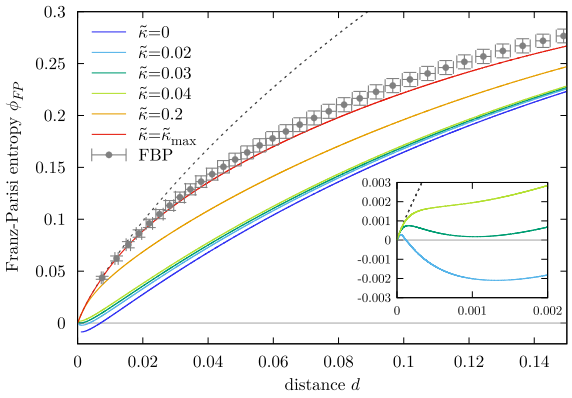

FIG. 3. Local entropy profiles (with $\kappa = 0$) of typical solutions at $\alpha = 0.5$ as a function of the distance, for various values of $\tilde{\kappa}$. The dashed line is the geometric upper bound obtained by counting all the configurations. The inset shows a detail of the three curves for $\tilde{\kappa} = 0.02$, 0.03, and 0.04 in the small-d range. We observe that for $\tilde{\kappa} = 0$ the solutions are isolated (the curve is missing for small distances due to numerical issues, but see [28]). For $0 < \tilde{\kappa} < \tilde{\kappa}_{\min} \simeq 0.03$ the entropy has a small positive dense region at small d and there is an interval where it is negative (see the $\tilde{\kappa} = 0.02$ curve). For $\tilde{\kappa}_{\min} \leq \tilde{\kappa} < \tilde{\kappa}_u \simeq 0.04$ the profiles are all positive, but there is a local maximum (see the $\tilde{\kappa} = 0.03$ curve). For larger $\tilde{\kappa}$, they grow monotonically up to the global maximum located at a large distance $d^*(\tilde{\kappa})$ (not visible). The entropy is a monotonic function of $\tilde{\kappa}$ for all distances up to $d^*(\tilde{\kappa}_{\max}) \simeq 0.285$, and the highest curve is the one for $\tilde{\kappa}_{\max} \simeq 0.418$. The points with error bars show the results of numerical experiments (ten samples at $N = 2001$ obtained with the focusing-BP (FBP) algorithm, which by design seeks high-local-entropy solutions; local entropy estimated by belief propagation, see SM [23]).

meaning that typical solutions with nonzero margin are always surrounded by an exponential number of solutions having zero margin. Furthermore, for small distances the local entropy is nearly indistinguishable from the geometric upper bound: almost all configurations around the reference solution are themselves solutions, up to a small, but still $O(N)$, Hamming distance. This means that the cluster is dense. (3) There exists a $\tilde{\kappa}_{\min}(\alpha) > 0$ such that if $0 < \tilde{\kappa} < \tilde{\kappa}_{\min}(\alpha)$ the local entropy is negative in an interval of distances $d \in [d_1, d_2]$ not containing the origin. This means that no solutions can be found in a spherical shell of radius $d \in [d_1, d_2]$. (4) There exists a $\tilde{\kappa}_u(\alpha) > 0$ such that if $\tilde{\kappa}_{\min}(\alpha) < \tilde{\kappa} < \tilde{\kappa}_u(\alpha)$ the local entropy is positive, but it is nonmonotonic. Notice that for $0 < \tilde{\kappa} < \tilde{\kappa}_u$ the local entropy develops a secondary maximum at short distances. This means that typical solutions with such $\tilde{\kappa}$ are immersed within small regions that have a characteristic size—they can be described as isolated (for $\tilde{\kappa} < \tilde{\kappa}_{\min}$) or "entropically" isolated (for $\tilde{\kappa} > \tilde{\kappa}_{\min}$) balls. (5) When $\tilde{\kappa} > \tilde{\kappa}_u(\alpha)$ [which can only happen if $\tilde{\kappa}_u(\alpha) < \tilde{\kappa}_{\max}(\alpha)$] the local entropy is monotonic up to the global maximum, at large distances. This suggests that typical solutions with large enough $\tilde{\kappa}$ are immersed in dense regions that do not seem to have a characteristic size and may extend to very large scales: the high-local-entropy regions. We speculate that this property is related to the accessibility of such regions by algorithms.

The picture described above stays qualitatively the same if we take $\kappa > 0$ and $\tilde{\kappa} \geq \kappa$. In particular, it is interesting to note that typical solutions with a given margin $\tilde{\kappa}$ are isolated with respect to solutions with the same margin $\kappa = \tilde{\kappa}$. However typical solutions with margin $\tilde{\kappa}$ are always surrounded by an exponential number of solutions with lower margin $\kappa < \tilde{\kappa}$. We conclude that even though the high-margin solutions are completely isolated from each other, they tend to be closer to and concentrated in the rare regions of high local entropy of lower margin solutions. These regions can then be seen as the union of typical isolated configurations that have a nonzero margin; these are in turn surrounded by solutions with smaller and smaller margin $\kappa < \tilde{\kappa}$.

Dense cluster threshold.—It has been previously discussed, using a large-deviation approach, how the geometrical structure of the high-local-entropy cluster changes with the number of patterns αN [15,34]. It was found that the cluster fractures above a certain value α_u. Numerical experiments show that this geometrical transition strongly affects the behavior of algorithms: α_u is conjectured to be an upper bound for the capacity of efficient learning of algorithms [19].

As discussed in point 5 above, a similar situation occurs when considering typical high-margin solutions. Let us define the value α'_u as the largest α for which the "large-scale" phase exists. It is characterized by the property $\tilde{\kappa}_u(\alpha'_u) = \tilde{\kappa}_{\max}(\alpha'_u)$. Beyond this value, only the "isolated balls" phase (points 3 and 4 in the previous section) remains. Indeed, we found this α'_u to be only slightly smaller than the upper bound α_u derived from the large-deviation analysis. Thus, α'_u can be used to provide an easier estimate for the algorithmic upper bound.

This is illustrated in Fig. 4, where we show some plots of $\phi_{\mathrm{FP}}[d; \alpha, \tilde{\kappa}_{\max}(\alpha), \kappa]$, and its derivative with respect to the

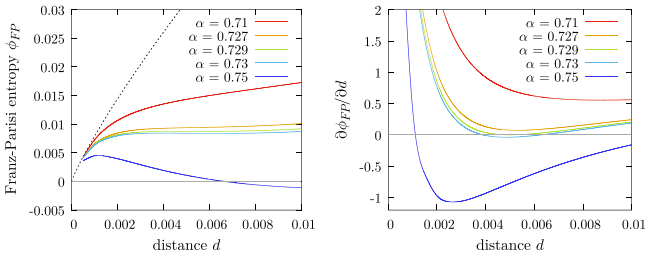

FIG. 4. Local entropy profiles (with $\kappa = 0$) of typical maximum margin solutions (left panel) and its derivative (right panel) as a function of the distance, for different values of α. For $\alpha = 0.71$ and 0.727 the entropy is monotonic, i.e., it has a unique maximum at large distances (not visible). For $\alpha = \alpha'_u \simeq 0.729$ the local entropy starts to be nonmonotonic (its derivative with respect to the distance develops a new zero). The entropy becomes negative at larger α [i.e., $\tilde{\kappa}_{\max}(\alpha) < \tilde{\kappa}_{\min}(\alpha)$] in a given range of distances.

distance, for several values of α. At $\alpha = \alpha'_u \simeq 0.73$ the derivative develops a new zero. In this case $\alpha_u \simeq 0.77$.

The discrepancy between the two thresholds can be mainly ascribed to the fact that in the derivation of α'_u only typical (albeit high margin) solutions are considered. On the other hand, the fact that $\alpha_u \approx \alpha'_u$ suggests that maximally dense solutions are not too dissimilar and not too far from maximum-margin solutions. To test this, we performed numerical experiments by sampling solutions found with the focusing-BP algorithm [19], which by design seeks maximally dense solutions, and measured their average local entropy using belief propagation (see SM [23] for details). We found that its local entropy profile is only slightly higher than that of the typical $\tilde{\kappa}_{\max}$ solutions, as shown in Fig. 3. This also agrees with previous findings concerning the distribution of stabilities of wide and flat minimizers [17] and the impact of certain losses, such as the cross-entropy [16], which induce a certain degree of robustness during training.

The fracturing transition that sets in when the curves become nonmonotonic is a complex phenomenon. It was first observed in the aforementioned large-deviations analysis as a transition in α. The current scheme allows us to detect the same transition by observing the space around typical solutions. In addition, we can also observe a transition in $\tilde{\kappa}$, where it intersects the value $\tilde{\kappa}_u(\alpha)$, and a transition in κ for fixed $\tilde{\kappa} > \tilde{\kappa}_u(\alpha)$. These transitions can be understood as the appearance of a characteristic distance identified by an entropic barrier beyond which the solutions sparsify dramatically.

Discussion and conclusions.—We have shown that the dense clusters of solutions which are accessed by algorithms in a nonconvex model of neural network coincide with regions of the weight space where high-margin solutions coalesce. While in these regions solutions with the same margin remain mutually isolated, they are connected through solutions of a smaller margin. These results shed light on accessibility and generalization properties, and hopefully can help in developing rigorous mathematical results for nonconvex neural networks. We have verified that similar phenomena take place in one-hidden-layer neural networks with binary and continuous weights (in the latter case, also with rectified linear unit activation functions; SM [23] Sec. III) and that numerical results on deeper networks corroborate the scenario (see SM [23] Sec. IV). Also, we refer to [38] for an analysis on a model with a nontrivial correlated pattern structure, which shows similar qualitative phenomena.

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

[2] D. W. Otter, J. R. Medina, and J. K. Kalita, IEEE Trans. Neural Networks Learn. Syst. **32**, 604 (2021).

[3] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, Nature (London) **577**, 706 (2020).

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, K. Tunyasuvunakool, O. Ronneberger, R. Bates, A. Žídek, A. Bridgland *et al.*, Nature (London) **596**, 583 (2021).

[5] P. Mehta, M. Bukov, C.-H. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, Phys. Rep. **810**, 1 (2019).

[6] L. Bottou, in *Proceedings of COMPSTAT'2010* (Springer, New York, 2010), pp. 177–186.

[7] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, arXiv:1609.04836.

[8] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, arXiv:1912.02178.

[9] G. K. Dziugaite and D. M. Roy, Entropy-SGD optimizes the prior of a PAC-bayes bound: Data-dependent PAC-bayes priors via differential privacy, arXiv:1712.09376.

[10] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht, 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, Stockholm, 2018), pp. 1309–1318.

[11] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, in *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., Montréal, 2018), Vol. 31.

[12] W. R. Huang, Z. Emam, M. Goldblum, L. Fowl, J. K. Terry, F. Huang, and T. Goldstein, in Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops, *Proceedings of Machine Learning Research Vol. 137*, edited by J. Zosa Forde, F. Ruiz, M. F. Pradier, and A. Schein (PMLR, Vancouver, 2020), pp. 87–97.

[13] P. C. Verpoort, A. A. Lee, and D. J. Wales, Proc. Natl. Acad. Sci. U.S.A. **117**, 21857 (2020).

[14] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Commun. ACM **64**, 107 (2021).

[15] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Phys. Rev. Lett. **115**, 128101 (2015).

[16] C. Baldassi, F. Pittorino, and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **117**, 161 (2020).

[17] C. Baldassi, E. M. Malatesta, and R. Zecchina, Phys. Rev. Lett. **123**, 170602 (2019).

[18] W. Zou and H. Huang, Phys. Rev. Research **3**, 033290 (2021).

[19] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **113**, E7655 (2016).

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

[21] J. Ding and N. Sun, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, Phoenix, 2019), pp. 816–827.

[22] E. Gardner and B. Derrida, J. Phys. A **22**, 1983 (1989).

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.127.278301 for the details of the analytical computations and numerical experiments, which includes Ref. [24].

[24] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, Binarized Neural Networks, *Advances in neural information processing Systems* 29, edited by D. Lee,

M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., Barcelona, 2016).
[25] E. Gardner, J. Phys. A **21**, 257 (1988).
[26] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[27] W. Krauth and M. Mézard, J. Phys. **50**, 3057 (1989).
[28] H. Huang and Y. Kabashima, Phys. Rev. E **90**, 052813 (2014).
[29] J. A. Zavatone-Veth and C. Pehlevan, Phys. Rev. E **103**, L020301 (2021).
[30] W. Perkins and C. Xu, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (Association for Computing Machinery, Virtual, 2021), pp. 1579–1588.
[31] E. Abbe, S. Li, and A. Sly, arXiv:2102.13069.
[32] A. Braunstein and R. Zecchina, Phys. Rev. Lett. **96**, 030201 (2006).
[33] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina, Proc. Natl. Acad. Sci. U.S.A. **104**, 11079 (2007).
[34] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, J. Stat. Mech. (2016) P023301.
[35] C. Baldassi, E. M. Malatesta, M. Negri, and R. Zecchina, J. Stat. Mech. (2020) 124012.
[36] Y. Feng and Y. Tu, Proc. Natl. Acad. Sci. U.S.A. **118**, e2015617118 (2021).
[37] S. Franz and G. Parisi, J. Phys. I **5**, 1401 (1995).
[38] C. Baldassi, C. Lauditi, E. M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, arXiv:2110.00683.