Properties of the Geometry of Solutions and Capacity of Multilayer Neural Networks with Rectified Linear Unit Activations

Carlo Baldassi, Enrico M. Malatesta,* and Riccardo Zecchina
Artificial Intelligence Lab, Institute for Data Science and Analytics, Bocconi University, Milano 20135, Italy

(Received 17 July 2019; published 23 October 2019)

Rectified linear units (ReLUs) have become the main model for the neural units in current deep learning systems. This choice was originally suggested as a way to compensate for the so-called vanishing gradient problem which can undercut stochastic gradient descent learning in networks composed of multiple layers. Here we provide analytical results on the effects of ReLUs on the capacity and on the geometrical landscape of the solution space in two-layer neural networks with either binary or real-valued weights. We study the problem of storing an extensive number of random patterns and find that, quite unexpectedly, the capacity of the network remains finite as the number of neurons in the hidden layer increases, at odds with the case of threshold units in which the capacity diverges. Possibly more important, a large deviation approach allows us to find that the geometrical landscape of the solution space has a peculiar structure: While the majority of solutions are close in distance but still isolated, there exist rare regions of solutions which are much more dense than the similar ones in the case of threshold units. These solutions are robust to perturbations of the weights and can tolerate large perturbations of the inputs. The analytical results are corroborated by numerical findings.

DOI: 10.1103/PhysRevLett.123.170602

Artificial neural networks (ANNs) have been studied for decades and yet only recently have they started to reveal their potentialities in performing different types of massive learning tasks [1]. Their current denomination is deep neural networks (DNNs) in reference to the choice of the architectures, which typically involve multiple interconnected layers of neuronal units. Learning in ANNs is in principle a very difficult optimization problem, in which "good" minima of the learning loss function in the high dimensional space of the connection weights need to be found. Luckily enough, DNN models have evolved rapidly, overcoming some of the computational barriers that for many years have limited their efficiency. Important components of this evolution have been the availability of computational power and the stockpiling of extremely rich data sets.

The features on which the various modeling strategies have intersected, besides the architectures, are the choice of the loss functions, the transfer functions for the neural units, and the regularization techniques. These improvements have been found to help the convergence of the learning processes, typically based on stochastic gradient descent (SGD) [2], and to lead to solutions which can often avoid overfitting even in overparametrized regimes. All these results pose basic conceptual questions which need to find a clear explanation in term of the optimization landscape.

Here we study the effects that the choice of the rectified linear units (ReLUs) for the neurons [3] has on the geometrical structure of the learning landscape. The ReLU is one of the most popular nonlinear activation functions and it has been extensively used to train DNNs, since it is known to dramatically reduce the training time for a typical algorithm [4]. It is also known that another major benefit of using the ReLU is that it is not as severely affected by the vanishing gradient problem as other transfer functions (e.g., tanh) [4,5]. We study ANN models with one hidden layer storing random patterns, for which we derive analytical results that are corroborated by numerical findings. At variance with what happens in the case of threshold units, we find that models built on ReLU functions present a critical capacity, i.e., the maximum number of patterns per weight which can be learned, that does not diverge as the number of neurons in the hidden unit is increased. At the same time we find that below the critical capacity they also present wider dense regions of solutions. These regions are defined in terms of the volume of the weights around a minimizer which do not lead to an increase of the loss value (e.g., number of errors) [6]. For discrete weights this notion reduces to the so-called local entropy [7] of a minimizer. We also check analytically and numerically the improvement in the robustness of these solutions with respect to both weight and input perturbations.

Together with the recent results on the existence of such wide flat minima and on the effect of choosing particular loss functions to drive the learning processes toward them [8], our result contributes to create a unified framework for the learning theory in DNN, which relies on the large

deviations geometrical features of the accessible solutions in the overparametrized regime.

The model.—We will consider a two-layer neural network with N input units, K neurons in the hidden layer, and one output. The mapping from the input to the hidden layer is realized by K nonoverlapping perceptrons each having N/K weights. Given $p = \alpha N$ inputs ξ^μ labeled by index $\mu = \{1, \ldots, p\}$, the output of the network for each input μ is computed as

$$\sigma^\mu_{\text{out}} = \text{sgn}\left(\frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l \tau^\mu_l\right) = \text{sgn}\left(\frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l g(\lambda^\mu_l)\right), \quad (1)$$

where λ^μ_l is the input of the l hidden unit, i.e., $\lambda^\mu_l = \sqrt{(K/N)}\sum_{i=1}^{N/K} W_{li}\xi^\mu_{li}$ and W_{li} is the weight connecting the input unit i to the hidden unit l. c_l is the weight connecting hidden unit l with the output; g is a generic activation function. In the following we will mainly consider two particular choices of activation functions and of the weights c_l. In the first one we take the sign activation $g(\lambda) = \text{sgn}(\lambda)$ and we fix to 1 all the weights c_l (in general their sign can be absorbed into the weights W_{li}). The $K = 1$ version of this model is the well-known perceptron and it has been extensively studied since the 1980s by means of the replica and cavity methods [9–11] used in spin glass theory [12]. The $K > 1$ case is known as the tree-committee machine that has been studied in the 1990s [13,14]. In the second model we will use the ReLU activation function that is defined with $g(\lambda) = \max(0, \lambda)$, and since the output of this transfer function is always non-negative we will fix half of the weights c_l to $+1$ and the remaining half to -1. Given a training set composed by random i.i.d. patterns $\xi^\mu \in \{-1, 1\}^N$ and labels $\sigma^\mu \in \{-1, 1\}$ and defining $\mathbb{X}_{\xi,\sigma}(W) \equiv \prod_\mu \theta[(\sigma^\mu/\sqrt{K})\sum_{l=1}^{K} c_l \tau^\mu_l]$, the weights that correctly classify the patterns are those for which $\mathbb{X}_{\xi,\sigma}(W) = 1$. Their volume (or number) is therefore [9,10]

$$Z = \int d\mu(W)\mathbb{X}_{\xi,\sigma}(W), \quad (2)$$

where $d\mu(W)$ is a measure over the weights W. In this study two constraints over the weights will be considered. The spherical constraint where for every $l \in \{1, \ldots, K\}$, we have $\sum_i W^2_{li} = N/K$, i.e., every subperceptron has weights that live on the hypersphere of radius $\sqrt{N/K}$. The second constraint we will use is the binary one, where for every $l \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, N/K\}$ we have $W_{li} \in \{-1, 1\}$. We are interested in the large K limit for which we will be able to compute analytically the capacity of the model for different choices of transfer function, to study the typical distances between absolute minima and to perform the large deviation study giving the local volumes associated to the wider flat minima.

Critical capacity.—We will analyze the problem in the limit of a large number N of input units. The standard scenario in this limit is that there is a sharp threshold α_c such that for $\alpha < \alpha_c$ the probability of finding a solution is 1 while for $\alpha > \alpha_c$ the volume of compatible weights is empty. α_c is therefore called *critical capacity* since it is the maximum number of patterns per weight that one can store in a neural network. The critical capacity of the mode, for a generic choice of the transfer function, can be evaluated computing the free entropy $\mathcal{F} \equiv (1/N)\langle \ln Z \rangle_{\xi,\sigma}$, where $\langle \cdots \rangle_{\xi,\sigma}$ denotes the average over the patterns, using the replica method; one finds

$$\mathcal{F} = \mathcal{G}_S + \alpha \mathcal{G}_E. \quad (3)$$

$\mathcal{G}_S$ is the entropic term, which represents the logarithm of the volume at $\alpha = 0$, where there are no constraints induced by the training set; this quantity is independent of K and it is affected only by the binary or spherical nature of the weights. $\mathcal{G}_E$ is the energetic term and it represents the logarithm of the fraction of solutions. Moreover it depends on the order parameters $q^{ab}_l \equiv (K/N)\sum_i W^a_{li} W^b_{li}$ which represent the overlap between subperceptrons l of two different replicas a and b of the machine. Using a replica-symmetric (RS) ansatz, in which we assume $q^{ab}_l = q$ for all a, b, l, and in the large K limit, $\mathcal{G}_E$ is

$$\mathcal{G}_E = \int Dz_0 \ln H\left(-\sqrt{\frac{\Delta - \Delta_{-1}}{\Delta_2 - \Delta}} z_0\right) \quad (4)$$

where $Dz \equiv (dz/\sqrt{2\pi})e^{-z^2/2}$ and $H(x) \equiv \int_x^\infty Dz$. This expression is equivalent to that of the perceptron (i.e., $K = 1$), the only difference being that the order parameters are replaced by effective ones that depend on the general activation function used in the machine [13]. In Eq. (4) we have called these effective order parameters Δ_{-1}, Δ, and Δ_2, see the Supplemental Material (SM) [15] for details, and in the perceptron they are 0, q, and 1, respectively.

In the binary case the critical capacity is always smaller than 1 and it is identified with the point where the RS free entropy $\mathcal{F}$ vanishes. This condition requires

$$\alpha_c = \frac{\frac{\hat{q}}{2}(1 - q) - \int Du \ln(2\cosh(\sqrt{\hat{q}}u))}{\int Dz_0 \ln H(-\sqrt{\frac{\Delta - \Delta_{-1}}{\Delta_2 - \Delta}} z_0)} \quad (5)$$

where $\hat{q}$ is the conjugated parameter of q. q and $\hat{q}$ are found by solving their associated saddle point equations (details in the SM [15]). Solving these equations one finds for the ReLU case $\alpha_c = 0.9039(9)$ which is a smaller value than in the sign activation function case, where one gets $\alpha_c = 0.9469(5)$ as shown in Ref. [13].

In the spherical case the situation is different since the capacity is not bounded from above. Previous works [13,14]

have shown, in the case of the sign activations, that the RS estimate of the critical capacity diverges with the number of neurons in the hidden layer as $\alpha_c \simeq (72K/\pi)^{1/2}$, violating the Mitchison-Durbin bound [16]. The reason for this discrepancy is due to the fact that the Gardner volume disconnects before α_c and therefore replica-symmetry breaking (RSB) takes place. Indeed, the instability of the RS solution occurs at a finite value $\alpha_{\rm AT} \simeq 2.988$ at large K. A subsequent work [17] based on multifractal techniques derived the correct scaling of the capacity with K as $\alpha_c \simeq (16/\pi)\sqrt{\ln K}$, which saturates the Mitchison-Durbin bound.

In the case of the ReLU functions the RS estimate of the critical capacity is obtained simply performing the $q \to 1$ limit, as for the perceptron. Quite surprisingly, if the activation function is such that $\Delta_2 - \Delta \simeq \delta\Delta(1-q)$ for $q \to 1$, with $\delta\Delta$ a finite proportionality term, the RS estimate of the critical capacity is finite (analogously to the case of the perceptron, where having exactly $\Delta_2 - \Delta = 1 - q$) makes the capacity finite. Contrary to the sign activation (where the effective parameters are such that $\Delta_2 - \Delta \simeq \delta\Delta\sqrt{1-q}$), the ReLU activation function happens to belong to this class (with $\delta\Delta = \frac{1}{2}$). The RS estimate of the critical capacity is therefore given by

$$\alpha_c^{\rm RS} = \frac{2\delta\Delta}{\Delta_2 - \Delta_{-1}}. \tag{6}$$

One correctly recovers $\alpha_c = 2$ in the case of the perceptron [9] whereas for the committee machine with ReLU activation one has $\alpha_c = 2[1 - (1/\pi)]^{-1} \simeq 2.934$. As for the sign activation, one expects also for the ReLU activation that the RS saddle point is unstable before α_c. Indeed we have computed the stability of the RS solution in the large K limit and we found $\alpha_{\rm AT} \simeq 0.615$ which is far smaller than the corresponding value of the sign activation. This suggests that strong RSB effects are at play.

We have therefore used a 1RSB ansatz to better estimate the critical capacity in the ReLU case. This can be obtained by taking the limits $q_1 \to 1$ for intrablock overlap and $m \to 0$ for the Parisi parameter. We found $\alpha_c^{\rm 1RSB} \simeq 2.92$, which is not too far from the RS result.

For the sake of brevity, we just mention that the results on the nondivergent capacity with K generalize to other monotone smooth functions such as the sigmoid.

Typical distances.—In order to get a quantitative understanding of the geometrical structure of the weight space, we have also derived the so called Franz-Parisi entropy for the committee machine with a generic transfer function. This framework was originally introduced in Ref. [18] to study the metastable states of mean-field spin glasses and only recently it was used to study the landscape of the solutions of the perceptron [19]. The basic idea is to sample a solution $\tilde{W}$ from the equilibrium Boltzmann measure and to study the entropy landscape around it. In the binary setting, it turns out that the equilibrium solutions of the learning problem are isolated; this means that, for any positive value of α, one must flip an extensive number of weights to go from an equilibrium solution to another one. The Franz-Parisi entropy is defined as

$$\mathcal{F}_{\rm FP}(S) = \frac{1}{N}\left\langle \frac{\int d\mu(\tilde{W})\mathbb{X}_{\xi,\sigma}(\tilde{W}) \ln \mathcal{N}_{\xi,\sigma}(\tilde{W}, S)}{\int d\mu(\tilde{W})\mathbb{X}_{\xi,\sigma}(\tilde{W})} \right\rangle_{\xi,\sigma} \tag{7}$$

where $\mathcal{N}_{\xi,\sigma}(\tilde{W}, S) = \int d\mu(W)\mathbb{X}_{\xi,\sigma}(W) \prod_{l=1}^{K} \delta[W_l \cdot \tilde{W}_l - (N/K)S]$ counts the number of solutions at a distance $d = (1 - S/2)$ from a reference $\tilde{W}$. The distance constraint is imposed by fixing the overlap between every subperceptron of W and $\tilde{W}$ to (S/K). The quantity defined in Eq. (7) can again be computed by the replica method. However, the expression for K finite is quite difficult to analyze numerically since the energetic term contains $4K$ integrals. The large K limit is instead easier and, again, the only difference with the perceptron expression is that the order parameters are replaced with effective ones in the energetic term (see SM [15] for details).

In Fig. 1 we plot the Franz-Parisi entropy $\mathcal{F}_{\rm FP}$ as a function of the distance $d = (1 - S/2)$ from a typical reference solution $\tilde{W}$ for the committee machine with both sign and ReLU activations. The numerical analysis shows that, as in the case of the binary perceptron, also in the 2-layer case solutions are also isolated since there is a minimal distance d^* below which the entropy becomes negative. This minimal distance increases with the constraint density α. However at a given value of α, typical solutions of the committee machine with ReLU activations are less isolated than the ones of the sign counterpart. The same framework applies to the case of spherical weights where we find that the minimum distance between typical solutions is smaller for the ReLU case.

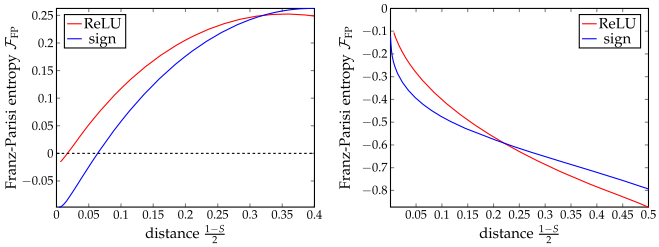

FIG. 1. (Left panel) Typical solutions are isolated in 2 layer neural networks with binary weights. We plot the Franz-Parisi entropy $\mathcal{F}_{\rm FP}$ as a function of the distance from a typical reference solution $\tilde{W}$ for the committee machine with ReLU activations (red line) and sign activations (blue line) in the limit of a large number of neurons in the hidden layer K. We have used $\alpha = 0.6$. For both curves there is a value of the distance for which the entropy becomes negative. This signals that typical solutions are isolated. (Right panel) Franz-Parisi entropy as a function of distance, normalized with respect to the unconstrained $\alpha = 0$ case, for spherical weights and for $\alpha = 1$.

Large deviation analysis.—The results of the previous section show that the Franz-Parisi framework does not capture the features of high local entropy regions. These regions indeed exist since algorithms can be observed to find solutions belonging to large connected clusters of solutions. In order to study the properties of wide flat minima or high local entropy regions one needs to introduce a large deviation measure [7], which favors configurations surrounded by an exponential number of solutions at small distance. This amounts to studying a system with y real replicas constrained to be at a distance d from a reference configuration $\tilde{W}$. The high local entropy region is found around $d \simeq 0$ in the limit of large y. As shown in Ref. [8], an alternative approach can be obtained by directly constraining the set of y replicas to be at a given mutual distance d, that is

$$Z_{\mathrm{LD}}(d,y) = \int \prod_{a=1}^{y} d\mu(W^a) \prod_{a=1}^{y} \mathbb{X}_{\xi,\sigma}(W^a) \times \prod_{\substack{a<b \\ l}} \delta\left(W_l^a \cdot W_l^b - \frac{N(1-2d)}{K} \right). \quad (8)$$

This last approach has the advantage of simplifying the calculations, since it is related to the standard 1RSB approach on the typical Gardner volume [20] given in Eq. (2): the only difference is that the Parisi parameter m and the intrablock overlap q_1 are fixed as external parameters, and play the same role of y and $1-2d$, respectively. Therefore m is not limited anymore to the standard range [0, 1]; indeed, the large m regime is the significant one for capturing high local entropy regions. In the large m and K limit the large deviation free entropy $\mathcal{F}_{\mathrm{LD}} \equiv (1/N)\langle \ln Z_{\mathrm{LD}} \rangle_{\xi,\sigma}$ reads

$$\mathcal{F}_{\mathrm{LD}}(q_1) = \mathcal{G}_S(q_1) + \alpha \mathcal{G}_E(q_1) \quad (9)$$

where, again, the entropic term has a different expression depending on the constraint over the weights W. Its expression, together with the corresponding energetic term, is reported in the SM [15].

We report in Fig. 2 the numerical results for both binary and spherical weights of the large deviation entropy (normalized with respect to the unconstrained $\alpha = 0$ case) as a function of q_1. For both sign and ReLU activations, the region $q_1 \simeq 1$ is flat around zero. This means that there exist $\tilde{W}$ references around which the landscape of solutions is basically indistinguishable from the $\alpha = 0$ case where all configurations are solutions. We also find that the ReLU curve, in the vicinity of $q_1 \simeq 1$, is always more entropic than the corresponding one of the sign. This picture is valid for sufficiently low values of α; for α greater of a certain value α^* the two curves switch. This is due to the fact that the two models have completely different critical capacities

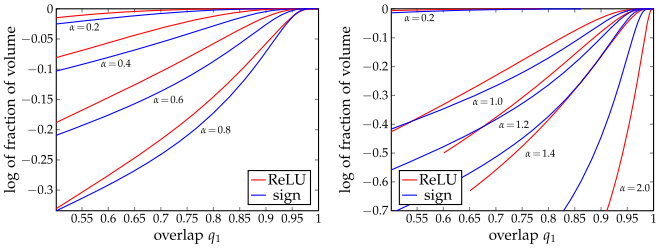

FIG. 2. Numerical evidence of the greater robustness of the minima of ReLU transfer function (red line) compared with the sign one (blue line) using the large deviation analysis in the binary (left panel) setting and spherical setting (right panel). The exchange in the curves in both settings for sufficiently large α is due to the fact that the algorithmic threshold of the ReLU is reached before than the corresponding one of the sign case.

(divergent in the sign case, finite in the ReLU case) so that one expects that clusters of solutions disappear at a smaller constrained capacity when ReLU activations are used.

Stability distribution and robustness.—To corroborate our previous results, we have also computed (details in the SM [15]), for various models and types of solutions W, the distribution of the *stabilities* $\Xi = \langle (\sigma/\sqrt{K}) \sum_{l=1}^{K} c_l \tau_l^{\mu} \rangle_{\xi,\sigma}$, which measure the distance from the threshold at the output unit in the direction of the correct label σ, cf. Eq. (1). Previous calculations [21] have shown that in the simple case of the spherical perceptron at the critical capacity the stability distribution around a typical solution W develops a delta peak in the origin $\Xi \simeq 0$; we confirmed that even in the two-layer case the stability distribution of a typical solution, being isolated, also has its mode at $\Xi \simeq 0$ even at lower α, see the dashed lines in Fig. 3 (left). A solution surrounded by an exponential number of other solutions, instead, should be more robust and be centered away from 0. Our calculations show that this is indeed the case both for the sign and for the ReLU activations, and we have confirmed the results by numerical simulations. In Fig. 3 (left) we show the analytical and numerical results for the

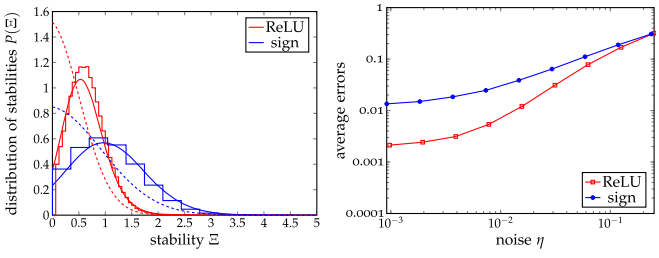

FIG. 3. (Left panel) Dashed lines: Theoretical stability curves for the typical solutions, for binary weights at $\alpha = 0.4$. Solid lines: Comparison between the numerical and theoretical stability distributions in the large deviation scenario, same α. We have used $q_1 = 0.85$, $y = 20$. (Right panel) Robustness of the reference configuration found by replicated simulated annealing when one pattern is perturbed by flipping a certain fraction η of entries.

case of binary weights at $\alpha = 0.4$ with $y = 20$ replicas at $q_1 = 0.85$. For the numerical results, we have used simulated annealing on a system with $K = 32$ ($K = 33$) for the ReLU (sign) activations (respectively), and $N = K^2 \simeq 10^3$. We have simulated a system of y interacting replicas that is able to sample from the local-entropic measure [6] with the RRR Monte Carlo method [22], ensuring that the annealing process was sufficiently slow such that at the end of the simulation all replicas were solutions, and controlling the interaction such that the average overlap between replicas was equal to q_1 within a tolerance of 0.01. The results were averaged over 20 samples. As seen in Fig. 3 (left), the agreement with the analytical computations is remarkable, despite the small values of N/K and K and the approximations introduced by sampling with simulated annealing.

The stabilities for the sign and ReLU activations are qualitatively similar, but quantitatively we observe that in all cases the curves for the ReLU case have a peak closer to 0 and a smaller variance. These are not, however, directly comparable, and it is difficult to tell from the stability curves alone which choice is more robust. We have thus directly measured, on the results of the simulations, the effect of introducing noise in the input patterns. For each trained group of y replicas, we used the configuration of the reference $\tilde{W}$ (which lays in the middle of the cluster of solutions) and we measured the probability that a pattern of the training set would be misclassified if perturbed by flipping a fraction η of randomly chosen entries. We explored a wide range of values of the noise η and sampled 50 perturbations per pattern. The results are shown in Fig. 3 (right), and they confirm that the networks with ReLU activations are more robust than those with sign activations for this α, in agreement with the results of Fig. 2. We also verified that the reference configuration is indeed more robust than the individual replicas. The results for other choices of the parameters are qualitatively identical. Our preliminary tests show that the same phenomenology is maintained when the network architecture is changed to a fully-connected scheme, in which each hidden unit is connected to all of the input units.

The architecture of the model that we have analyzed here is certainly very simplified compared to state-of-the-art deep neural networks used in applications. Investigating deeper models would certainly be of great interest, but extremely challenging with current analytical techniques, and is thus an open problem. Extending our analysis to a one-hidden-layer fully-connected model, on the other hand, would in principle be feasible (the additional complication comes from the permutation symmetry of the hidden layer). However, based on the existing literature (e.g., Refs. [13,23]), and our preliminary numerical experiments mentioned above, we do not expect that such an extension would result in qualitatively different outcomes compared to our treelike model.

C. B. and R. Z. acknowledge the ONR Grant No. N00014-17-1-2569.

*enrico.malatesta@unibocconi.it

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

[2] L. Bottou, Large-scale machine learning with stochastic gradient descent, in *Proceedings of COMPSTAT'2010* (Springer, Heidelberg, 2010), pp. 177–186, https://doi.org/10.1007/978-3-7908-2604-3_16.

[3] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit, Nature (London) **405**, 947 (2000).

[4] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, edited by G. Gordon, D. Dunson, and M. Dudík, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (PMLR, Fort Lauderdale, FL, USA 2011), Vol. 15, pp. 315–323.

[5] S. Hochreiter, Untersuchungen zu dynamischen neuronalen netzen, Diploma, Technische Universität München, 91 (1991).

[6] 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).

[7] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant Dense Clusters Allow for Simple Learning and High Computational Performance in Neural Networks with Discrete Synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[8] C. Baldassi, F. Pittorino, and R. Zecchina, Shaping the learning landscape in neural networks around wide flat minima, arXiv:1905.07833.

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

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

[11] M. Mézard, The space of interactions in neural networks: Gardner's computation with the cavity method, J. Phys. A **22**, 2181 (1989).

[12] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, International Series of Monographs on Physics (Oxford University Press, Oxford, 2001).

[13] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Phys. Rev. A **45**, 4146 (1992).

[14] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Storage capacity and learning algorithms for two-layer neural networks, Phys. Rev. A **45**, 7590 (1992).

[15] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.123.170602 for the details of the calculations.

[16] G. Mitchison and R. Durbin, Bounds on the learning capacity of some multi-layer networks, Biol. Cybern. **60**, 345 (1989).

[17] R. Monasson and R. Zecchina, Weight Space Structure and Internal Representations: A Direct Approach to Learning and Generalization in Multilayer Neural Networks, Phys. Rev. Lett. **75**, 2432 (1995).

[18] S. Franz and G. Parisi, Recipes for metastable states in spin glasses, J. Phys. I **5**, 1401 (1995).
[19] H. Huang and Y. Kabashima, Origin of the computational hardness for learning with binary synapses, Phys. Rev. E **90**, 052813 (2014).
[20] W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, J. Phys. (France) **50**, 3057 (1989).
[21] T. B Kepler and L. F Abbott, Domains of attraction in neural networks, J. Phys. (France) **49**, 1657 (1988).
[22] C. Baldassi, A method to reduce the rejection rate in Monte Carlo Markov chains. J. Stat. Mech. (2017) 033301.
[23] R Urbanczik, Storage capacity of the fully-connected committee machine., J. Phys. A **30**, L387 (1997).