Properties of the geometry of solutions and capacity of multi-layer neural networks with Rectified Linear Units activations

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

Rectified Linear Units (ReLU) 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 (SGD) 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.

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 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 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 over parametrized 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 (ReLU) for the neurons [3] has on the geometrical structure of the learning landscape. ReLU is one of the most popular non-linear 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. the 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 non-overlapping perceptrons each having N/K weights. Given $p = \alpha N$ inputs ξ^μ labeled by index $\mu = \{1, \dots, p\}$, the output of the network for each input μ is computed as

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

where λ^μ_l is the input of the l hidden unit i.e. $\lambda^\mu_l = \sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi^\mu_{li}$ 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 '80s 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 '90s [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 $\mathbb{X}_{\xi,\sigma}(W) \equiv \prod_\mu \theta\left(\frac{\sigma^\mu}{\sqrt{K}} \sum_{l=1}^K c_l \tau_l^\mu\right)$, 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]

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

where $\mathrm{d}\mu(W)$ is a measure over the weights W. In this study two constraints over the weights will be considered. The spherical constraint where for every $l \in \{1, \dots, K\}$, we have $\sum_i W_{li}^2 = N/K$, i.e. every sub-perceptron 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 $l \in \{1, \dots, K\}$ and $i \in \{1, \dots, N/K\}$ 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 $\mathcal{F} \equiv \frac{1}{N} \langle \ln Z \rangle_{\xi,\sigma}$, where $\langle \cdot \rangle_{\xi,\sigma}$ denotes the average over the patterns, using the replica method; one finds

$$\mathcal{F} = \mathcal{G}_S + \alpha\, \mathcal{G}_E\,. \tag{3}$$

$\mathcal{G}_S$ 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 $q_l^{ab} \equiv \frac{K}{N} \sum_i W_{li}^a W_{li}^b$ which represent the overlap between sub-perceptrons l of two different replicas a and b of the machine. Using a replica-symmetric (RS) ansatz, in which we assume $q_l^{ab} = q$ for all a, b, l, and in the large K limit, $\mathcal{G}_E$ is

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

where $Dz \equiv \frac{\mathrm{d}z}{\sqrt{2\pi}} e^{-z^2/2}$ 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 as Δ_{-1}, Δ, and Δ_2, see the appendix B 1 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

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

where $\hat{q}$ is the conjugated parameter of q. q and $\hat{q}$ are found by solving their associated saddle point equations (details in appendix B 1). 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 $\alpha_c \simeq \left(\frac{72K}{\pi}\right)^{1/2}$, violating the Mitchison-Durbin bound [15]. 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 $\alpha_{\mathrm{AT}} \simeq 2.988$ at large K. A subsequent work [16] based on multifractal techniques derived the correct scaling of the capacity with K as $\alpha_c \simeq \frac{16}{\pi}\sqrt{\ln K}$, 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

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

One correctly recovers $\alpha_c = 2$ in the case of the perceptron [9] whereas for the committee machine with ReLU activation one has $\alpha_c = 2\left(1 - \frac{1}{\pi}\right)^{-1} \simeq 2.934$. 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 $\alpha_{\mathrm{AT}} \simeq 0.615$

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 intra-block overlap and $m \to 0$ for the Parisi parameter. We found $\alpha_c^{\text{1RSB}} \simeq 2.92$, which is not too far from the RS result.

For the sake of brevity, we just mention that the results on the non divergent 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. [17] 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 [18]. 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

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

where $\mathcal{N}_{\xi,\sigma}\left(\tilde{W}, S\right) = \int \mathrm{d}\mu(W)\, \mathbb{X}_{\xi,\sigma}(W) \prod_{l=1}^{K} \delta\left(W_l \cdot \tilde{W}_l - \frac{N}{K} S\right)$ counts the number of solutions at a distance $d = \frac{1-S}{2}$ from a reference $\tilde{W}$. The distance constraint is imposed by fixing the overlap between every sub-perceptron of W and $\tilde{W}$ to $\frac{S}{K}$. 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 appendix C for details).

In Fig. 1 we plot the Franz-Parisi entropy $\mathcal{F}_{\text{FP}}$ as a function of the distance $d = \frac{1-S}{2}$ 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, 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.

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

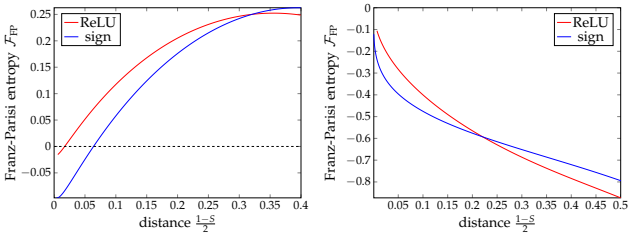

Figure 1. (Left panel) Typical solutions are isolated in 2 layer neural networks with binary weights. We plot the Franz-Parisi entropy $\mathcal{F}_{\text{FP}}$ as a function of the distance from a typical reference solution $\tilde{W}$ for the committee machine with ReLU activations (red line) and sign 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 $\alpha = 1$.

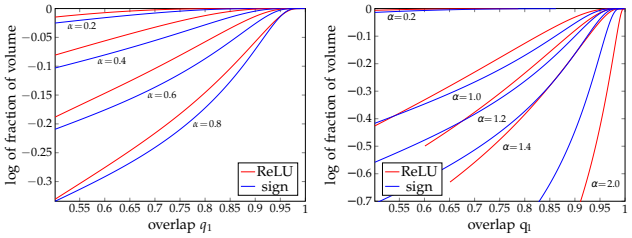

Figure 2. Numerical evidence of the greater robustness of the minima of ReLU transfer function (red line) compared with the sign one (blue line) using the large deviation analysis in the binary (left panel) setting and spherical setting (right panel). The exchange in the curves in both settings for sufficiently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

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

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

This last approach has the advantage of simplifying the calculations, since it is related to the standard 1RSB approach on the typical Gardner volume [19] given in eq. (2): the only difference is that the Parisi parameter m and the intra-block

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 $\mathcal{F}_{\rm LD} \equiv \frac{1}{N}\langle \ln Z_{\rm LD}\rangle_{\xi,\sigma}$ reads

$$\mathcal{F}_{\rm LD}(q_1) = \mathcal{G}_S(q_1) + \alpha\,\mathcal{G}_E(q_1) \tag{9}$$

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 appendix D.

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 (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 (see appendix E), for various models and types of solutions W, the distribution of the *stabilities* $\Xi = \left\langle \frac{\sigma}{\sqrt{K}} \sum_{l=1}^{K} c_l\, \tau_l^{\mu} \right\rangle_{\xi,\sigma}$, which measure the distance from the threshold at the output unit in the direction of the correct label σ, cf. eq. (1). Previous calculations [20] 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 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 [21], 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

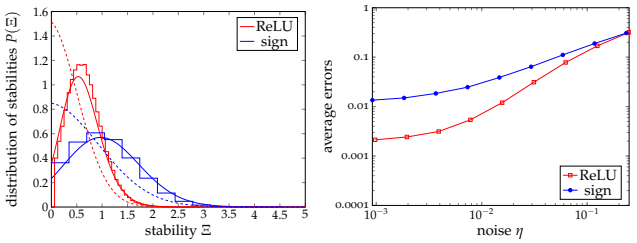

Figure 3. (Left panel) Dashed lines: theoretical stability curves for the typical solutions, for binary weights at $\alpha = 0.4$. Solid lines: comparison between the numerical and theoretical stability distributions in the large deviation scenario, same α. We have used $q_1 = 0.85$, $y = 20$. (Right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

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, 22]), and our preliminary numerical experiments mentioned above, we do not expect

that such an extension would result in qualitatively different outcomes compared to our tree-like model.

Acknowledgments. C.B. and R.Z. acknowledge the ONR grant N00014-17-1-2569.

* enrico.malatesta@unibocconi.it

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi:10.1038/nature14539.

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. doi:10.1007/978-3-7908-2604-3_16.

[3] Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947, 2000. doi:10.1038/35016072.

[4] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL: http://proceedings.mlr.press/v15/glorot11a.html.

[5] Sepp Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991.

[6] Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proceedings of the National Academy of Sciences*, 113(48):E7655–E7662, November 2016. doi:10.1073/pnas.1608103113.

[7] Carlo Baldassi, Alessandro Ingrosso, Carlo Lucibello, Luca Saglietti, and Riccardo Zecchina. Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Phys. Rev. Lett.*, 115:128101, Sep 2015. doi:10.1103/PhysRevLett.115.128101.

[8] Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *arXiv preprint arXiv:1905.07833*, 2019. URL: https://arxiv.org/abs/1905.07833.

[9] Elizabeth Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257–270, jan 1988. doi:10.1088/0305-4470/21/1/030.

[10] Elizabeth Gardner and Bernard Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and General*, 21(1):271–284, jan 1988. doi:10.1088/0305-4470/21/1/031.

[11] Marc Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181, 1989. doi:doi.org/10.1088/0305-4470/22/12/018.

[12] Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. International series of monographs on physics. Oxford University Press, 2001. doi:10.1093/acprof:oso/9780198509417.001.0001.

[13] Eli Barkai, David Hansel, and Haim Sompolinsky. Broken symmetries in multilayered perceptrons. *Phys. Rev. A*, 45:4146–4161, Mar 1992. doi:10.1103/PhysRevA.45.4146.

[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*, 45:7590–7609, May 1992. doi:10.1103/PhysRevA.45.7590.

[15] Graeme Mitchison and Richard Durbin. Bounds on the learning capacity of some multi-layer networks. *Biological Cybernetics*, 60(5):345–365, 1989. doi:10.1007/BF00204772.

[16] Rémi Monasson and Riccardo Zecchina. Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks. *Physical review letters*, 75(12):2432, 1995. doi:10.1103/PhysRevLett.75.2432.

[17] Silvio Franz and Giorgio Parisi. Recipes for metastable states in spin glasses. *Journal de Physique I*, 5(11):1401–1415, 1995. doi:10.1051/jp1:1995201.

[18] Haiping Huang and Yoshiyuki Kabashima. Origin of the computational hardness for learning with binary synapses. *Physical Review E*, 90(5):052813, 2014. doi:https://doi.org/10.1103/PhysRevE.90.052813.

[19] Werner Krauth and Marc Mézard. Storage capacity of memory networks with binary couplings. *J. Phys. France*, 50:3057–3066, 1989. doi:10.1051/jphys:0198900500200305700.

[20] Thomas B Kepler and Laurence F Abbott. Domains of attraction in neural networks. *J. Phys. France*, 49(10):1657–1662, 1988. doi:10.1051/jphys:0198800490100165700.

[21] Carlo Baldassi. A method to reduce the rejection rate in monte carlo markov chains. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(3):033301, 2017. doi:10.1088/1742-5468/aa5335.

[22] R Urbanczik. Storage capacity of the fully-connected committee machine. *Journal of Physics A: Mathematical and General*, 30(11):L387, 1997. doi:10.1088/0305-4470/30/11/007.

Appendix A: Model

Our model is a tree-like committee machine with N weights W_{li} divided into K groups of N/K entries. We use the index $l = 1, \dots, K$ for the group and $i = 1, \dots, \frac{N}{K}$ for the entry. We consider two cases, the binary case $W_{li} = \pm 1$for all l, i and the continuous case with spherical constraints on each group, $\sum_{i=1}^{N/K} W_{li}^2 = N/K$ for all l.

The training set consists of $p = \alpha N$ random binary i.i.d. patterns. The inputs are denoted by ξ_{li}^{μ} and the outputs by σ^{μ}, where $\mu = 1, \dots, p$ is the pattern index.

In our analysis, we write the model using a generic activation function $g(x)$ for the first layer (hidden) units. We will consider two cases: the sign case $g(x) = \mathrm{sgn}(x)$ and the ReLU case $g(x) = \max(0, x)$. The output of any given unit l in response to a

pattern μ is thus written as

$$\tau_l^\mu = g\left(\frac{1}{\sqrt{N/K}}\sum_i W_{li}\xi_{li}^\mu\right) \tag{A1}$$

The connection weights between the first layer and the output are denoted by c_l and considered binary and fixed; for the case of ReLU activations we set the first half $l = 1, \dots, K/2$ to the value $+1$ and the rest to -1; for the case of the sign activations we can set them to all to $+1$ without loss of generality.

A configuration of the weights W solves the training problem if it classifies correctly all the patterns; we denote this with the indicator function

$$\mathbb{X}_{\xi,\sigma}(W) = \prod_\mu \Theta\left(\frac{\sigma^\mu}{\sqrt{K}}\sum_l c_l\,\tau_l^\mu\right) \tag{A2}$$

where $\Theta(x) = 1$ if $x > 0$ and 0 otherwise is the Heaviside step function.

The volume of the space of configurations that correctly classify the whole training set is then

$$Z = \int \mathrm{d}\mu(W)\,\mathbb{X}_{\xi,\sigma}(W) \tag{A3}$$

where $\mathrm{d}\mu(W)$ is the flat measure over the admissible values of the W, depending on whether we're analyzing the binary or the spherical case. The average of the log-volume over the distribution of the patterns, $\langle \log Z\rangle_{\xi,\sigma}$, is the free entropy of the model $\mathcal{F}$. We can evaluate it in the large N limit with the the "replica trick", using the formula $\log Z = \lim_{n\to 0}\partial_n Z^n$, computing the average for integer n and then taking the limit $n \to 0$.

As explained in the main text, in all cases the resulting expression takes the form

$$\mathcal{F} = \mathcal{G}_S + \alpha\mathcal{G}_E \tag{A4}$$

where the $\mathcal{G}_S$ part is only affected by the spherical or binary nature of the weights, whereas the $\mathcal{G}_E$ part is only affected by K and by the activation function g. Determining their value requires to compute a saddle-point over some overlap parameters q_l^{ab} with $a, b = 1, \dots, n$ representing overlaps between replicas, and their conjugates $\hat{q}_l^{ab}$; in turn, this requires an ansatz about the structure of the saddle-point in order to perform the $n \to 0$ limit.

For the spherical weights, the $\mathcal{G}_S$ part (before the $n \to 0$ limit) reads

$$\mathcal{G}_S^{\text{sph}} = -\frac{1}{2nK}\sum_l\sum_{a,b} q_l^{ab}\hat{q}_l^{ab} + \frac{1}{nK}\sum_l \log\int\prod_a \mathrm{d}W^a\; e^{\frac{1}{2}\sum_{a,b}\hat{q}_l^{ab}W^aW^b} \tag{A5}$$

while for the binary case we have a very similar expression, except that the summations don't have the $a = b$ case and the integral over W^a becomes a summation:

$$\mathcal{G}_S^{\text{bin}} = -\frac{1}{2nK}\sum_{l=1}^K\sum_{a\neq b} q_l^{ab}\hat{q}_l^{ab} + \frac{1}{nK}\sum_{l=1}^K \log\sum_{W^a=\pm 1} e^{\frac{1}{2}\sum_{a\neq b}\hat{q}_l^{ab}W^aW^b} \tag{A6}$$

The $\mathcal{G}_E$ part reads:

$$\mathcal{G}_E = \frac{1}{n}\mathbb{E}_\sigma \log\int\prod_{la}\frac{\mathrm{d}u_l^a\mathrm{d}\hat{u}_l^a}{2\pi}\prod_a\Theta\left(\sigma^\mu\frac{1}{\sqrt{K}}\sum_l c_l g\left(u_l^a\right)\right)e^{-\sum_{al} i\hat{u}_l^a u_l^a - \frac{1}{2}\sum_l\sum_{ab} q_l^{ab}\hat{u}_l^a\hat{u}_l^b} \tag{A7}$$

In all cases, we study the problem in the large K limit, which allows to invoke the central limit theorem and leads to a crucial simplification of the expressions.

Appendix B: Critical capacity

1. Replica symmetric ansatz

In the replica-symmetric (RS) case we seek solutions of the form $q_l^{ab} = \delta_{ab} + (1-\delta_{ab})\,q$ for all l, a, b, where δ_{ab} is the Kronecker delta symbol, and similarly for the conjugated parameters, $\hat{q}_l^{ab} = \delta_{ab}\hat{Q} + (1-\delta_{ab})\,\hat{q}$ for all l, a, b. The resulting

expressions, as reported in the main text, are:

$$\mathcal{G}_S^{\text{sph}} = \frac{1}{2}\hat{Q} + \frac{1}{2}q\hat{q} + \frac{1}{2}\ln\frac{2\pi}{\hat{Q}+\hat{q}} + \frac{\hat{q}}{2\left(\hat{Q}+\hat{q}\right)} \tag{B1}$$

$$\mathcal{G}_S^{\text{bin}} = -\frac{\hat{q}}{2}(1-q) + \int Dz\, \ln 2\cosh\left(z\sqrt{\hat{q}}\right) \tag{B2}$$

$$\mathcal{G}_E = \int Dz_0\, \ln H\left(-\sqrt{\frac{\Delta-\Delta_{-1}}{\Delta_2-\Delta}}z_0\right) \tag{B3}$$

where $Dz \equiv \mathrm{d}z\frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}$ is a Gaussian measure, $H(x) = \int_x^\infty Dz = \frac{1}{2}\mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right)$ and the expressions of Δ, Δ_2 and Δ_{-1} depend on the activation function g:

$$\Delta_{-1}^{\text{sgn}} = 0; \qquad \Delta^{\text{sgn}} = 1 - \frac{2}{\pi}\arccos(q); \qquad \Delta_2^{\text{sgn}} = 1$$

$$\Delta_{-1}^{\text{ReLU}} = \frac{1}{2\pi}; \qquad \Delta^{\text{ReLU}} = \frac{\sqrt{1-q^2}}{2\pi} + \frac{q}{\pi}\arctan\sqrt{\frac{1+q}{1-q}}; \qquad \Delta_2^{\text{ReLU}} = \frac{1}{2}$$

The values of the overlaps and conjugated parameters are found by setting to 0 the derivatives of the free entropy.

The critical capacity α_c is found in the binary case by seeking numerically the value of α for which the saddle point solutions returns a zero free entropy.For the spherical case, instead, α_c is determined by finding the value of α such that $q \to 1$, which can be obtained analytically by reparametrizing $q = 1 - \delta q$ and expanding around $\delta q \ll 1$. In this limit, we must also reparametrize Δ using $\Delta_2 - \Delta = \delta\Delta\, \delta q^x$, where x is an exponent that depends on the activation function: it is $x = 1/2$ for the sign and $x = 1$ for the ReLU. Due to this difference in this exponent, α_c diverges in the sign activation case (as was shown in [13, 14]), while for the ReLU activations it converges to $2\left(1-\frac{1}{\pi}\right)^{-1}$. However, the RS result for the spherical case is only an upper bound, and a more accurate result requires replica-symmetry breaking.

2. 1RSB ansatz

In the one-step replica-symmetry-breaking (1-RSB) ansatz we seek solutions with 3 possible values of the overlaps q_l^{ab} and their conjugates. We group the n replicas in $\frac{n}{m}$ groups of m replicas each, and denote with q_0 the overlaps among different groups and with q_1 the overlaps within the same group. As before, the self overlap is $q^{aa} = 1$ and its conjugate $\hat{q}^{aa} = \hat{Q}$ (these are only relevant in the spherical case).

The resulting expressions are:

$$\mathcal{G}_S^{\text{sph}} = \frac{1}{2}\hat{Q} + \frac{1}{2}\hat{q}_1 q_1 + \frac{m}{2}(q_0\hat{q}_0 - q_1\hat{q}_1) + \frac{1}{2}\left[\ln\frac{2\pi}{\hat{Q}+\hat{q}_1} + \frac{1}{m}\ln\left(\frac{\hat{Q}+\hat{q}_1}{\hat{Q}+\hat{q}_1 - m(\hat{q}_1-\hat{q}_0)}\right) + \frac{\hat{q}_0}{\hat{Q}+\hat{q}_1-m(\hat{q}_1-\hat{q}_0)}\right] \tag{B4}$$

$$\mathcal{G}_S^{\text{bin}} = -\frac{\hat{q}_1}{2}(1-q_1) + \frac{m}{2}(q_0\hat{q}_0 - q_1\hat{q}_1) + \frac{1}{m}\int Du\, \ln\int Dv\left(2\cosh\left(\sqrt{\hat{q}_0}u + \sqrt{\hat{q}_1-\hat{q}_0}v\right)\right)^m \tag{B5}$$

$$\mathcal{G}_E = \frac{1}{m}\int Dz_0\, \ln\int Dz_1 H\left(-\frac{\sqrt{\Delta_0-\Delta_{-1}}z_0 + \sqrt{\Delta_1-\Delta_0}z_1}{\sqrt{\Delta_2-\Delta_1}}\right)^m \tag{B6}$$

the expressions of Δ_2 and Δ_{-1} are the same as in the RS case. The expressions of Δ_0 and Δ_1 take the same form as the RS expressions for Δ, except that q_0 and q_1 must be used instead of q.

Similarly to the RS case, in order to determine the critical capacity α_c in the spherical case, we need to find the value of α such that $q_1 \to 1$. In this case however we must also have $m \to 0$. The scaling is such that $\tilde{m} = m/(1-q_1)$ is finite. The final expression reads

$$\alpha_c = \frac{\ln(1+\tilde{m}(1-q_0)) + \frac{\tilde{m}q_0}{1+\tilde{m}(1-q_0)}}{2f(q_0,\tilde{m})} \tag{B7}$$

where

$$f\left(q_0,\tilde{m}\right)=\int Dz_0\,\ln\left[\frac{\sqrt{\delta\Delta}\,e^{-\frac{(\Delta_0-\Delta_{-1})\tilde{m}}{\delta\Delta+(\Delta_2-\Delta_0)\tilde{m}}\frac{z_0^2}{2}}}{\sqrt{\delta\Delta+(\Delta_2-\Delta_0)\,\tilde{m}}}H\left(-\sqrt{\frac{\Delta_0-\Delta_{-1}}{\Delta_2-\Delta_0}}\frac{\sqrt{\delta\Delta}\,z_0}{\sqrt{\delta\Delta+(\Delta_2-\Delta_0)\,\tilde{m}}}\right)+H\left(\sqrt{\frac{\Delta_0-\Delta_{-1}}{\Delta_2-\Delta_0}}z_0\right)\right] \tag{B8}$$

The expression of $\delta\Delta$ is the same as in the RS case using Δ_1 instead of Δ; the values of q_0 and $\tilde{m}$ are determined by saddle point equations as usual.

Appendix C: Franz-Parisi potential

The Franz-Parisi entropy [17, 18] is defined as (cf. eq. (7) of the main text):

$$\mathcal{F}_{\rm FP}\left(S\right)=\frac{1}{N}\left\langle\frac{\int \mathrm{d}\mu\left(\tilde{W}\right)\,\mathbb{X}_{\xi,\sigma}\left(\tilde{W}\right)\ln\mathcal{N}_{\xi,\sigma}\left(\tilde{W},S\right)}{\int \mathrm{d}\mu\left(\tilde{W}\right)\,\mathbb{X}_{\xi,\sigma}\left(\tilde{W}\right)}\right\rangle_{\xi,\sigma} \tag{C1}$$

where $\mathcal{N}_{\xi,\sigma}\left(\tilde{W},S\right)=\int \mathrm{d}\mu(W)\,\mathbb{X}_{\xi,\sigma}(W)\prod_{l=1}^{K}\delta\left(W_l\cdot\tilde{W}_l-\frac{N}{K}S\right)$ counts the number of solutions at a distance $\frac{1-S}{2}$ from the reference $\tilde{W}$. In this expression, $\tilde{W}$ represents a typical solution. The evaluation of this quantity requires the use of two sets of replicas: n replicas of the reference configuration $\tilde{W}$, for which we use the indices a and b, and r replicas of the surrounding configurations W, for which we use the indices c and d. The computation proceeds following standard steps, leading to these expressions, written in terms of the overlaps $q_l^{ab}=\frac{K}{N}\sum_{i=1}^{N/K}\tilde{W}_{li}^a\tilde{W}_{li}^b$, $p_l^{cd}=\frac{K}{N}\sum_{i=1}^{N/K}W_{li}^cW_{li}^d$, $t_l^{ac}=\frac{K}{N}\sum_{i=1}^{N/K}\tilde{W}_{li}^aW_{li}^c$ and their conjugate quantities:

$$\mathcal{F}_{\rm FP}=\mathcal{G}_S+\alpha\,\mathcal{G}_E \tag{C2}$$

$$\begin{aligned}\mathcal{G}_S^{\rm sph}=&-\frac{1}{2nK}\sum_{l=1}^{K}\sum_{a,b}q_l^{ab}\hat{q}_l^{ab}-\frac{1}{2nK}\sum_{l=1}^{K}\sum_{c,d}p_l^{cd}\hat{p}_l^{cd}-\frac{1}{nK}\sum_{l=1}^{K}\sum_{a>1,c}t_l^{ac}\hat{t}_l^{ac}-\frac{S}{nK}\sum_{l=1}^{K}\sum_{a>1,c}\hat{S}_l^c+\\&+\frac{1}{nK}\ln\int\prod_{al}\mathrm{d}\tilde{W}_l^a\prod_{cl}\mathrm{d}W_l^c\prod_l e^{\frac{1}{2}\sum_{a,b}\hat{q}_l^{ab}\tilde{W}_l^a\tilde{W}_l^b+\frac{1}{2}\sum_{c,d}\hat{p}_l^{cd}W_l^cW_l^d+\sum_{a>1,c}\hat{t}_l^{ac}\tilde{W}_l^aW_l^c+\tilde{W}_l^{a=1}\sum_c\hat{S}_l^cW_l^c}\end{aligned} \tag{C3}$$

$$\begin{aligned}\mathcal{G}_S^{\rm bin}=&-\frac{1}{2nK}\sum_{l=1}^{K}\sum_{a\neq b}q_l^{ab}\hat{q}_l^{ab}-\frac{1}{2nK}\sum_{l=1}^{K}\sum_{c\neq d}p_l^{cd}\hat{p}_l^{cd}-\frac{1}{nK}\sum_{l=1}^{K}\sum_{a>1,c}t_l^{ac}\hat{t}_l^{ac}-\frac{S}{nK}\sum_{l=1}^{K}\sum_{a>1,c}\hat{S}_l^c+\\&+\frac{1}{nK}\ln\sum_{\tilde{W}_l^a=\pm1}\sum_{W_l^c=\pm1}\prod_l e^{\frac{1}{2}\sum_{a\neq b}\hat{q}_l^{ab}\tilde{W}_l^a\tilde{W}_l^b+\frac{1}{2}\sum_{c\neq d}\hat{p}_l^{cd}W_l^cW_l^d+\sum_{a>1,c}\hat{t}_l^{ac}\tilde{W}_l^aW_l^c+\tilde{W}_l^{a=1}\sum_c\hat{S}_l^cW_l^c}\end{aligned} \tag{C4}$$

$$\begin{aligned}\mathcal{G}_E=&\frac{1}{n}\mathbb{E}_\sigma\ln\int\prod_{la}\frac{\mathrm{d}\lambda_l^a\mathrm{d}\hat{\lambda}_l^a}{2\pi}\prod_{lc}\frac{\mathrm{d}u_l^c\mathrm{d}\hat{u}_l^c}{2\pi}e^{i\lambda_l^a\hat{\lambda}_l^a+iu_l^c\hat{u}_l^c}\prod_a\theta\left(\frac{\sigma}{\sqrt{K}}\sum_l c_l\,g\left(\lambda_l^a\right)\right)\prod_c\theta\left(\frac{\sigma}{\sqrt{K}}\sum_l c_l\,g\left(u_l^c\right)\right)\times\\&\times\prod_l e^{-\frac{1}{2}\sum_a\left(\hat{\lambda}_l^a\right)^2-\frac{1}{2}\sum_c\left(\hat{u}_l^c\right)^2-\sum_{a<b}q_l^{ab}\hat{\lambda}_l^a\hat{\lambda}_l^b-\sum_{c<d}p_l^{cd}\hat{u}_l^c\hat{u}_l^d-\sum_{a>1,c}t_l^{ac}\hat{\lambda}_l^a\hat{u}_l^c-S\hat{\lambda}_l^{a=1}\sum_c\hat{u}_l^c}\end{aligned} \tag{C5}$$

The calculation proceeds by taking the RS ansatz with the same structure as that of sec. B 1 and the limit $n\to0$, $r\to0$. We obtain, in the large K limit:

$$\mathcal{G}_S^{\rm sph} = \frac{1}{2}\hat{P} + \frac{1}{2}\hat{p}p + \hat{t}t - \hat{S}S + \frac{1}{2}\ln\frac{2\pi}{\hat{P}+\hat{p}} + \frac{1}{\hat{P}+\hat{p}}\left[\frac{\hat{p}}{2} + \frac{\left(\hat{S}-\hat{t}\right)^2\left(\frac{\hat{Q}}{2}+\hat{q}\right)}{\left(\hat{Q}+\hat{q}\right)^2} + \frac{\hat{t}\left(\hat{S}-\hat{t}\right)}{\hat{Q}+\hat{q}}\right] \tag{C6}$$

$$\mathcal{G}_S^{\rm bin} = -\frac{1}{2}\hat{p}\,(1-p) + \hat{t}t - \hat{S}S + \int Du \frac{\sum_{\tilde{W}=\pm1} e^{\tilde{W}\sqrt{\hat{q}}x}\int Dv \ln\left[2\cosh\left(\sqrt{\hat{p}-\frac{\hat{t}^2}{\hat{q}}}v + \frac{\hat{t}}{\sqrt{\hat{q}}}u + \left(\hat{S}-\hat{t}\right)\tilde{W}\right)\right]}{2\cosh\left(\sqrt{\hat{q}}u\right)} \tag{C7}$$

$$\mathcal{G}_E = \int Dz_0 \frac{\int Dz_1 H\left(-\frac{D_1 z_1 + \sqrt{\Sigma_0\Gamma}z_0}{\sqrt{\Sigma_1\Gamma - D_1^2}}\right)\ln H\left(-\frac{\sqrt{\Gamma}z_1 + D_0 z_0/\sqrt{\Sigma_0}}{\sqrt{\Delta_3}}\right)}{H\left(-\sqrt{\frac{\Sigma_0}{\Sigma_1}}z_0\right)} \tag{C8}$$

where we introduced the auxiliary quantities Σ_0, Σ_1, Γ, D_0, D_1 and Δ_3 which depend on the choice of the activation function (like the $\Delta_{-1/0/1/2}$ of the previous section). For the sign activations we get:

$$\Delta_3^{\rm sgn} = \frac{2}{\pi}\arccos p \tag{C9}$$

$$\Sigma_0^{\rm sgn} = 1 - \frac{2}{\pi}\arccos q \tag{C10}$$

$$\Sigma_1^{\rm sgn} = \frac{2}{\pi}\arccos q \tag{C11}$$

$$D_0^{\rm sgn} = \frac{2}{\pi}\arctan\left(\frac{t}{\sqrt{1-t^2}}\right) \tag{C12}$$

$$D_1^{\rm sgn} = \frac{2}{\pi}\left[\arctan\left(\frac{S}{\sqrt{1-S^2}}\right) - \arctan\left(\frac{t}{\sqrt{1-t^2}}\right)\right] \tag{C13}$$

$$\Gamma^{\rm sgn} = 1 - \frac{2}{\pi}\arccos p - \frac{\left(D_0^{\rm sgn}\right)^2}{\Sigma_0^{\rm sgn}} \tag{C14}$$

For the ReLU activations we get:

$$\Delta_3^{\rm ReLU} = \frac{1}{2} - \frac{\sqrt{1-p^2}}{2\pi} - \frac{p}{\pi}\arctan\sqrt{\frac{1+p}{1-p}} \tag{C15}$$

$$\Sigma_0^{\rm ReLU} = \frac{\sqrt{1-q^2}}{2\pi} + \frac{q}{\pi}\arctan\sqrt{\frac{1+q}{1-q}} - \frac{1}{2\pi} \tag{C16}$$

$$\Sigma_1^{\rm ReLU} = \frac{1}{2} - \frac{\sqrt{1-q^2}}{2\pi} - \frac{q}{\pi}\arctan\sqrt{\frac{1+q}{1-q}} \tag{C17}$$

$$D_0^{\rm ReLU} = \frac{\sqrt{1-t^2}}{2\pi} + \frac{t}{\pi}\arctan\sqrt{\frac{1+t}{1-t}} - \frac{1}{2\pi} \tag{C18}$$

$$D_1^{\rm ReLU} = \frac{\sqrt{1-S^2}}{2\pi} + \frac{S}{\pi}\arctan\sqrt{\frac{1+S}{1-S}} - \frac{\sqrt{1-t^2}}{2\pi} - \frac{t}{\pi}\arctan\sqrt{\frac{1+t}{1-t}} \tag{C19}$$

$$\Gamma^{\rm ReLU} = \frac{\sqrt{1-p^2}}{2\pi} + \frac{p}{\pi}\arctan\sqrt{\frac{1+p}{1-p}} - \frac{1}{2\pi} - \frac{\left(D_0^{\rm ReLU}\right)^2}{\Sigma_0^{\rm ReLU}} \tag{C20}$$

In order to find the order parameters for any given α and S, we need to set to 0 the derivatives of the free entropy w.r.t. the order parameters q, p, t and the conjugates $\hat{Q}$, $\hat{q}$, $\hat{P}$, $\hat{p}$, $\hat{t}$, $\hat{S}$, thus obtaining a system of 9 equations (7 for the binary case) to be solved numerically. The equations actually reduce to 6 (5 in the binary case) since q, $\hat{Q}$ and $\hat{q}$ are the same ones derived from the typical case (sec. B 1).

Appendix D: Large deviation analysis

Following [8], the large deviation analysis for the description of the high-local-entropy landscape uses the same equations as the standard 1RSB expressions eqs. (B4), (B5) and (B6). In this case, however, the overlap q_1 is not determined by a saddle point equation, but rather it is treated as an external parameter that controls the mutual overlap between the y replicas of the system. Also, the parameter m is not optimized and it is not restricted to the range $[0, 1]$; instead, it plays the role of the number of replicas y and it is generally taken to be large (we normally use either a large integer number to compare the results with numerical simulations, or we take the limit $m \to \infty$). For these reason, there are two saddle point equations less compared to the standard 1RSB calculation.

The resulting expression for the free entropy $\mathcal{F}_{\rm LD}(q_1)$ represents, in the spherical case, the log-volume of valid configurations (solutions at the correct overlap) of the system of y replicas. These configurations are thus embedded in $\mathcal{S}^{Ky}$ where $\mathcal{S}$ is the N/K-dimensional sphere of radius $\sqrt{N/K}$. In order to quantify the solution density, we must normalize $\mathcal{F}_{\rm LD}(q_1)$, subtracting the log-volume of all the admissible configurations at a given q_1 without the solution constraint (which is obtained by the analogous computation with $\alpha = 0$). The resulting quantity is thus upper-bounded by 0 (cf. Fig. (2) of the main text). For the binary case, $\mathcal{F}_{\rm LD}(q_1)$ is the log of the number of admissible solutions, and the same normalization procedure can be applied.

Large m limit

$$\mathcal{G}_S^{\rm bin}(q_1) = -\frac{\hat{q}_1}{2}(1-q_1) - \frac{1}{2}(\delta q_0\,\hat{q}_1 + q_1\delta\hat{q}_0) + \int Du \max_v \left[-\frac{v^2}{2} + \ln 2\cosh\left(\sqrt{\hat{q}_1}u + \sqrt{\delta\hat{q}_0}v\right)\right] \tag{D1}$$

$$\mathcal{G}_S^{\rm sph}(q_1) = \frac{1}{2}\frac{q_1}{1-q_1+\delta q_0} + \frac{1}{2}\ln(1-q_1) \tag{D2}$$

$$\mathcal{G}_E(q_1) = \int Dz_0 \max_{z_1}\left[-\frac{z_1^2}{2} + \log H\left(-\frac{\sqrt{\Delta_1-\Delta_{-1}}z_0 + \sqrt{\delta\Delta_0}\,z_1}{\sqrt{\Delta_2-\Delta_1}}\right)\right] \tag{D3}$$

In the $m \to \infty$ case the order parameters q_0 and $\hat{q}_0$ need to be rescaled with m and reparametrized with two new quantities δq_0 and $\delta\hat{q}_0$, as follows:

$$q_0 = q_1 - \frac{\delta q_0}{m} \tag{D4}$$

$$\hat{q}_0 = \hat{q}_1 - \frac{\delta\hat{q}_0}{m} \tag{D5}$$

As a consequence, we also reparametrize Δ_0 with a new parameter $\delta\Delta_0$ defined as:

$$m(\Delta_1 - \Delta_0) = \delta\Delta_0 = \begin{cases} \frac{2\delta q_0}{\pi\sqrt{1-q_1^2}} & \text{(sign)} \\ \frac{\delta q_0}{\pi}\arctan\left(\sqrt{\frac{1+q_1}{1-q_1}}\right) & \text{(ReLU)} \end{cases} \tag{D6}$$

The expressions eqs. (B4), (B5) and (B6) become:

Appendix E: Distribution of stabilities

The stability for a given pattern/label pair ξ^*, σ^* is defined as:

$$\Xi(\xi^*, \sigma^*) = \frac{\sigma^*}{\sqrt{K}}\sum_{l=1}^{K} c_l\, g\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi^*_{li}\right) \tag{E1}$$

The distribution over the training set for a typical solutions can thus be computed as

$$P(\Xi) = \left\langle \frac{\int \mathrm{d}\mu(W)\, \mathbb{X}_{\xi,\sigma}(W)\, \delta\left(\Xi - \Xi\left(\xi^1, \sigma^1\right)\right)}{\int \mathrm{d}\mu(W)\, \mathbb{X}_{\xi,\sigma}(W)} \right\rangle_{\xi,\sigma} \tag{E2}$$

where we arbitrarily chose the first pattern/label pair ξ^1, σ^1 without loss of generality. The expression can be computed by the replica method as usual, and the order parameters are simply obtained from the solutions of the saddle point equations for the free entropy. The resulting expression at the RS level is:

$$P(\Xi) = \Theta(\Xi) \int Dz \, \frac{G\left(\frac{\Xi - z\sqrt{\Delta - \Delta_{-1}}}{\sqrt{\Delta_2 - \Delta}}\right)}{\Delta_2 - \Delta} H\left(-z\sqrt{\frac{\Delta - \Delta_{-1}}{\Delta_2 - \Delta}}\right)^{-1} \tag{E3}$$

where $G(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is a standard Gaussian. The difference between the models (spherical/binary and sign/ReLU) is encoded in the different values for the overlaps and in the different expressions for the parameters Δ, Δ_{-1}, Δ_2.

In the large deviation case, we simply compute the expression with a 1RSB ansatz and fix q_1 and m as described in the previous section. The resulting expression is

$$P(\Xi) = \frac{\Theta(\Xi)}{\sqrt{\Delta_2 - \Delta_1}} \int Dz_0 \frac{\int Dz_1 G\left(\frac{\Xi - z_0\sqrt{\Delta_0 - \Delta_{-1}} + z_1\sqrt{\Delta_1 - \Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right) H\left(-\frac{z_0\sqrt{\Delta_0 - \Delta_{-1}} + z_1\sqrt{\Delta_1 - \Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right)^{m-1}}{\int Dz_1 H\left(-\frac{z_0\sqrt{\Delta_0 - \Delta_{-1}} + z_1\sqrt{\Delta_1 - \Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right)^{m}} \tag{E4}$$

where the effective parameters Δ_{-1} , Δ_0, Δ_1 and Δ_2 are the same defined in section B 2. In the $m \to \infty$ limit the previous expression reduces to

$$P(\Xi) = \Theta(\Xi) \int Dz_0 \frac{G\left(\frac{\Xi - z_0\sqrt{\Delta_1 - \Delta_{-1}} + z_1^*\sqrt{\delta\Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right)}{\sqrt{\Delta_2 - \Delta_1}} H\left(-\frac{z_0\sqrt{\Delta_1 - \Delta_{-1}} + z_1^*\sqrt{\delta\Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right)^{-1} \tag{E5}$$

where z_1^* satifies

$$z_1^* = \mathrm{argmax}_{z_1}\left[-\frac{z_1^2}{2} + \ln\left(-\frac{z_0\sqrt{\Delta_1 - \Delta_{-1}} + z_1\sqrt{\delta\Delta_0}}{\sqrt{\Delta_2 - \Delta_1}}\right)\right] \tag{E6}$$

where $\delta\Delta_0$ is defined in equation (D6).