Shaping the learning landscape in neural networks around wide flat minima – Supporting Information

Carlo Baldassi,[1,2] Fabrizio Pittorino,[3,1] and Riccardo Zecchina[1,4]

[1] *Artificial Intelligence Lab, Institute for Data Science and Analytics, Bocconi University, 20136 Milan, Italy*
[2] *Istituto Nazionale di Fisica Nucleare, Sezione di Torino, 10125 Turin, Italy*
[3] *DISAT, Politecnico di Torino, 10129 Turin, Italy*
[4] *International Centre for Theoretical Physics, 34151 Trieste, Italy*

CONTENTS

I. HIGH LOCAL ENTROPY STATES FROM THE 1-RSB FORMALISM

Given a system described by a vector of discrete variables W with an associated energy function $E(W)$, the Boltzmann equilibrium distribution at inverse temperature β reads

$$P(W;\beta) = \frac{1}{Z(\beta)} e^{-\beta E(W)} \tag{S1}$$

where the normalization factor Z is given by the partition function

$$Z(\beta) = \sum_{W} e^{-\beta E(W)} \tag{S2}$$

In the limit $\beta \to \infty$, the distribution is just a flat measure over the ground states of the energy; we can denote the ground state energy as $E^\star = \min_W E(W)$ and the characteristic function over the ground states as

$$\mathbb{X}(W) = \begin{cases} 1 & \text{if } E(W) = E^\star \\ 0 & \text{otherwise} \end{cases} \tag{S3}$$

such that $Z(\infty) = \sum_W \mathbb{X}(W)$ and $\log Z(\infty)$ gives the entropy of the ground states.

In ref. [1], we introduced a large-deviation measure with a modified energy function in which each configuration is reweighted by a "local entropy" term. There, we only considered the $\beta \to \infty$ limit and defined the local entropy as the number of ground states at a certain normalized distance D from a reference configuration $\tilde{W}$:

$$\mathcal{S}\left(\tilde{W}, d\right) = \log \mathcal{N}\left(\tilde{W}, d\right) = \log \sum_{W} \mathbb{X}(W)\, \delta\left(d\left(W, \tilde{W}\right) - ND\right) \tag{S4}$$

where $d(\cdot,\cdot)$ is a suitably defined distance function and $\delta(\cdot)$ is the Kronecker delta. With this definition, we can define a modified partition function as follows:

$$Z(\infty, y, D) = \lim_{\beta\to\infty} \sum_{\tilde{W}} e^{-\beta E(\tilde{W}) + y\mathcal{S}(\tilde{W}, D)} \tag{S5}$$

which (up to irrelevant constant factors) coincides with:

$$Z(\infty, y, D) = \sum_{\tilde{W}} \mathbb{X}\left(\tilde{W}\right) \mathcal{N}\left(\tilde{W}, D\right)^y \tag{S6}$$

This approach can be shown to be strictly related to the 1-step replica-symmetry-broken (1RSB) formalism of the bare energetic problem. First, let us define a local free entropy at any given inverse temperature β' (i.e. a generalization of eq. (S4)), and use a soft-constraint on the distance through a Lagrange multiplier λ:

$$\phi\left(\tilde{W}, \lambda, \beta'\right) = \log \sum_{W} e^{-\beta' E(W) - \frac{\lambda}{2} d(W, \tilde{W})} \tag{S7}$$

Note that the use of a Lagrange multiplier is mostly convenient in order to make the relation with the 1RSB description evident. It is only equivalent to using a hard constraint for the distance in the thermodynamic limit, and depending on the convexity properties of the function ϕ, but we will generally ignore these issues for the time being and come back to them later.

We can then rewrite the large-deviation partition function eq. (S5) in this more general case as:

$$Z(\beta, \beta', y, \lambda) = \sum_{\tilde{W}} e^{-\beta E(\tilde{W}) + y\phi(\tilde{W}, \lambda, \beta')} \tag{S8}$$

Let us now consider the case in which $y \in \mathbb{N}$: this allows us, by simple algebraic manipulations, to rewrite the partition function introducing a sum over all the configurations of y replicas of the system:

$$Z(\beta, \beta', y, \lambda) = \sum_{\tilde{W}, \{W^a\}} e^{-\beta E(\tilde{W}) - \beta' \sum_{a=1}^{y} E(W^a) - \frac{\lambda}{2} \sum_{a=1}^{y} d(W^a, \tilde{W})} \tag{S9}$$

This partition function describes a system of $y+1$ interacting *real* replicas with an interaction that is mediated by the reference configuration $\tilde{W}$. However we can isolate the sum over the configurations of $\tilde{W}$ to obtain a system of y interacting real replicas. In the special case $\beta = 0$ we obtain:

$$Z(0, \beta', y, \lambda) = \sum_{\tilde{W}, \{W^a\}} e^{-\beta' \sum_{a=1}^{y} E(W^a) + \log \sum_{\tilde{W}} \exp\left(-\frac{\lambda}{2} \sum_{a=1}^{y} d(W^a, \tilde{W})\right)} \tag{S10}$$

We have stressed the fact that the replicas are real to avoid the confusion with the *virtual* replicas used for the "replica trick": here, we are not replicating the system virtually in order to compute a free entropy in the limit of zero replicas: instead, we are describing a system of y identical interacting objects. The general case of $y \in \mathbb{R}$ can be obtained by analytic continuation once an expression for all integer y is found.

This description is highly reminiscent of – in fact, almost identical to – the derivation of the ergodicity-breaking scheme used in ref. [2]: there, an auxiliary symmetry breaking field is introduced (having the same role of $\tilde{W}$ in our notation); then, a free energy expression is introduced in which the role of the energy is taken by a "local free entropy" (the analogous of eq. (S4) for general β), after which the system is replicated y times and the auxiliary field $\tilde{W}$ is traced out, leading to a system of y real replicas with an effective pairwise interaction. Finally, the limit of vanishing interaction ($\lambda \to 0^+$) is taken in order to derive the equilibrium description. When this system is studied in the replica-symmetric (RS) Ansatz, it results in the 1RSB description of the original system, with y having the role of the Parisi parameter (usually denoted by m). Indeed, in this limit of vanishing interaction and for $\beta = 0$, our equations reduce to the 1RSB case as well.

Therefore, apart from minor differences, the main point of discrepancy between that analysis and our approach is that we don't restrict ourselves to the equilibrium distribution. Instead, we explore the whole range of values of λ. In this context, we also have no reason to restrict ourselves to the range $y \in [0, 1]$, as it is usually done in order to give a physical interpretation to the 1RSB solution; to the contrary, we are (mostly) interested in the limit of large y, in which only the configurations of maximal local entropy are described.

The relationship between our analysis and the usual 1RSB case can be made even more direct, leading to an alternative – although with very similar results – large deviations analysis: consider, instead of eq. (S10), a partition function in which the interaction among the replicas is pairwise (without the reference configuration $\tilde{W}$) and the constraint on the distance is hard (introduced via a Dirac delta function):

$$Z_{1RSB}\left(\beta', y, D\right) = \sum_{\{W^a\}} e^{-\beta' \sum_{a=1}^{y} E(W^a)} \prod_{a>b} \delta\left(d\left(W^a, W^b\right) - ND\right) \tag{S11}$$

Suppose then that we study the average free entropy $\langle \log Z_{1RSB}\left(\beta', y, \gamma\right)\rangle$ (where $\langle\cdot\rangle$ represents the average over the quenched parameters, if any) in the context of replica theory. Then, we will have n virtual replicas of the whole system, and since each system has y real replicas we end up with ny total replicas. Let's use indices c, d for the virtual replicas and a, b for the real ones, such that a configuration will now have two indices, e.g. W^{ca}. Suppose that we manage to manipulate the expression such that it becomes a function, among other order parameters, of the overlaps $q^{ca,db} = \frac{1}{N}\left\langle W^{ca}, W^{db}\right\rangle$, where $\langle\cdot,\cdot\rangle$ represents some inner product, and that the distance function $d\left(\cdot,\cdot\right)$ can be expressed in terms of those. Then, as usual, we would introduce auxiliary integrals

$$\int \prod_{(ca,db)} \left(N \mathrm{d}q^{ca,db}\right) \prod_{(ca,db)} \delta\left(N q^{ca,db} - \left\langle W^{ca}, W^{db}\right\rangle\right) \tag{S12}$$

Using this, we can rewrite the interaction term. Say that $d\left(W, W'\right) = \langle W, W\rangle + \langle W', W'\rangle - 2\langle W, W'\rangle$, then:

$$\prod_{c}\prod_{a>b} \delta\left(d\left(W^{ca}, W^{cb}\right) - ND\right) = \prod_{c}\prod_{a>b} \delta\left(N\left(q^{ca,ca} + q^{cb,cb} - 2q^{ca,cb} - D\right)\right) \tag{S13}$$

By assuming replica symmetry, we seek a saddle point with this structure:

$$\begin{aligned} q^{ca,ca} &= Q \\ q^{ca,cb} &= q_1 \quad (a \neq b) \\ q^{ca,db} &= q_0 \quad (c \neq d) \end{aligned} \tag{S14}$$

with $Q \geq q_1 \geq q_0$. The interaction term eq. (S13) becomes:

$$\prod_{c}\prod_{a>b} \delta\left(N\left(q^{ca,ca} + q^{cb,cb} - 2q^{ca,cb} - D\right)\right) = \delta\left(2N\left(Q - q_1 - D\right)\right) \tag{S15}$$

Therefore, the external parameter D eliminates a degree of freedom in the solution to the saddle point equations for the overlaps. The final step in the replica calculation would have the form

$$\begin{aligned} \langle \log Z_{1RSB}\left(\beta', y, D\right)\rangle &= \phi_{1RSB}\left(\beta', y, Q, q_1, q_0, \dots\right) \delta\left(Q - q_1 - D\right) \\ &= \phi_{1RSB}\left(\beta', y, Q, Q - D, q_0, \dots\right) \end{aligned} \tag{S16}$$

where ϕ_{1RSB} is the expression that would have been derived in an equilibrium computation without the interaction term, the dots in the argument represent extra order parameters, and the order parameters are fixed by the saddle point equations

$$\begin{aligned} \partial_Q \phi_{1RSB}\left(\beta', y, Q, Q - D, q_0, \dots\right) &= 0 \\ \partial_{q_0} \phi_{1RSB}\left(\beta', y, Q, Q - D, q_0, \dots\right) &= 0 \\ &\vdots \end{aligned} \tag{S17}$$

Thus, the difference with respect to the usual 1RSB computation is that the equation for finding the extremum over q_1 is removed, and the one for finding the extremum over Q is modified. Maximizing over D, by solving for $\partial_D \phi = 0$, is then equivalent to the usual 1RSB description (equivalent to the case $\lambda \to 0$ in the soft-constraint case):

$$Z_{1RSB}\left(\beta', y\right) = \max_{D} Z_{1RSB}\left(\beta', y, D\right) \tag{S18}$$

In the common case where Q is fixed (e.g. if the variables W are discrete, or constraints on the norm are introduced) then this representation fixes q_1; it is clear then that our large deviations analysis (the alternative one of eq. (S11)) is simply derived by fixing q_1 as an external parameter, and thus omitting the saddle point equation $\partial_{q_1}\phi_{1RSB} = 0$. Note that this wouldn't make physical sense in the standard derivation of the 1RSB equations, since in that context q_1 is only introduced as an overlap between virtual replicas when choosing an Ansatz for the solutions of the saddle point equations; our derivation is only physically meaningful when describing a system of real interacting replicas or, in the case of the original derivation from eq. (S5), a system with a modified energy function.

II. CROSS-ENTROPY MINIMA, ERRORS AND HIGH LOCAL ENTROPY CONFIGURATIONS

A. Cross-Entropy loss ground states

In order to study analytically the properties of the minima of the CE loss function in the case of i.i.d. random patterns, the key obstacle is to compute the normalization factor of the Gibbs measure, the partition function Z. Once this is done one has access to the Gibbs measure, which concentrates on the minima of the loss in the $\beta \to \infty$ limit.

Z is an exponentially fluctuating random variable and in order to find its most probable values we need to average its logarithm, a complicated task which we perform by the replica method. Once this is done, the typical value of Z can be recovered by $Z_{\text{typ}} \simeq \exp\left(N \langle \log Z \rangle_\xi\right)$, where $\langle \cdot \rangle_\xi$ stands for the average over the random patterns.

We refer to ref. [3] for a thorough review of the replica method. Here we just remind the reader that the replica method is an analytic continuation technique which allows in some cases (mean-field models) to compute the expectation of the logarithm of the partition function from the knowledge of its integer moments. The starting point is the following small n expansion

$$Z^n = 1 + n \log Z + O\left(n^2\right)$$

This identity may be averaged over the random patterns and gives the average of the log from the averaged n-th power of the partition function

$$\langle \log Z \rangle_\xi = \min_{n \to 0} \frac{\langle Z^n \rangle_\xi - 1}{n}$$

The idea of the replica method is to restrict to integer n and to take the analytic continuation $n \to 0$

$$\langle Z^n \rangle_\xi = \prod_{a=1}^{n} \langle Z_a \rangle_\xi = \sum_{\{W^1, \dots, W^n\}} \left\langle e^{-\beta \sum_{a=1}^{n} E(W^a)} \right\rangle_\xi$$

We have n replicas of the initial model. The random patterns in the expression of the energy disappear once the average has been carried out. Eventually one computes the partition function of an effective system of nN variables with a non random energy function resulting from the average. The result may be written formally as

$$\langle Z^n \rangle_\xi = \exp\left(N F\left(n\right)\right)$$

where F is the expression resulting from the sum over all configurations. Once the small n limit is taken, the final expression can be estimated analytically by means of the saddle-point method given that N is assumed to be large.

In the case of our problem we have

$$Z = \sum_{\{w_i = \pm 1\}} \exp\left(-\beta \sum_{\mu=1}^{M} f\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\mu\right)\right)$$

Following the replica approach, we need to compute

$$\langle Z^n \rangle = \left\langle \int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \prod_{\mu,a} \exp\left(-\beta f\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^a \xi_i^\mu\right)\right)\right\rangle_\xi$$

where the integration measure is just over the binary values of the weights. By enforcing $x^\mu = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\mu$ through a delta function, we can linearize the dependence on the randomness of the patterns and perform the average as follows:

$$\langle Z^n\rangle = \left\langle \int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \int \prod_{a,\mu} \frac{\mathrm{d}x^{a\mu}\mathrm{d}\hat{x}^{a\mu}}{2\pi} \prod_{a,\mu} \exp\left(-\beta f\left(x^{a\mu}\right)\right) \prod_{a,\mu} \exp\left(i\hat{x}^{a\mu}x^{a\mu} - i\hat{x}^{a\mu}\sum_{i=1}^{N}\frac{w_i^a\xi_i^\mu}{\sqrt{N}}\right)\right\rangle_\xi =$$

$$= \int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \int \prod_{a,\mu} \frac{\mathrm{d}x^{a\mu}\mathrm{d}\hat{x}^{a\mu}}{2\pi} \prod_{a\mu} \exp\left(-\beta f\left(x^{a\mu}\right)\right) \prod_{a,\mu} \exp\left(i\hat{x}^{a\mu}x^{a\mu}\right) \exp\left(-\frac{1}{2N}\sum_{ab}\hat{x}^{a\mu}\hat{x}^{b\mu}\sum_i \frac{w_i^a w_i^b}{N}\right) =$$

$$= \int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \int \prod_{a,\mu} \frac{\mathrm{d}x^{a\mu}\mathrm{d}\hat{x}^{a\mu}}{2\pi} \int \prod_{a>b} \frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}}{2\pi} \prod_{a>b} e^{-Nq^{ab}\hat{q}^{ab}} \prod_{a>b} e^{\hat{q}^{ab}\sum_i w_i^a w_i^b} \prod_{a\mu} \exp\left(-\beta f\left(x^{a\mu}\right)\right) \times$$

$$\times \prod_{a,\mu} \exp\left(i\hat{x}^{a\mu}x^{a\mu}\right) \exp\left(-\frac{1}{2}\sum_{ab}\hat{x}^{a\mu}\hat{x}^{b\mu}q^{ab}\right) =$$

$$= \int \prod_{a>b} \frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}N}{2\pi} \prod_{a>b} e^{-Nq^{ab}\hat{q}^{ab}} \left(\int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \prod_{a>b} e^{\hat{q}^{ab}\sum_i w_i^a w_i^b}\right) \times$$

$$\times \left(\int \prod_a \frac{\mathrm{d}x^a\mathrm{d}\hat{x}^a}{2\pi} \prod_a \exp\left(-\beta\, f\left(x^a\right)\right) \prod_a \exp\left(i\hat{x}^a x^a\right) \exp\left(-\frac{1}{2}\sum_{ab}\hat{x}^a\hat{x}^b q^{ab}\right)\right)^{\alpha N}$$

where we have used the delta functions to introduce the order parameters q^{ab} and $\hat{q}^{ab}$. In order to write the multiple integrals in a form that can be evaluated by saddle point, we restrict to the replica symmetric assumption $q^{ab} = q$ and $\hat{q}^{qb} = \hat{q}$, and perform few simplifications.

First we sum over the weights:

$$\int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \prod_{a>b} e^{\hat{q}\sum_i w_i^a w_i^b} = \int \prod_{i,a} \mathrm{d}\mu\left(w_i^a\right) \prod_i e^{\frac{\hat{q}}{2}\left(\sum_a w_i^a\right)^2} e^{-\frac{\hat{q}}{2}Nn} =$$

$$= e^{-\frac{\hat{q}}{2}Nn} \left(\sum_{w=\pm 1} e^{\frac{\hat{q}}{2}\left(\sum_a w^a\right)^2}\right)^N = e^{-\frac{\hat{q}}{2}Nn} \left(\int Du \sum_{w=\pm 1} e^{\sqrt{\hat{q}}u\sum_a w^a}\right)^N =$$

$$\simeq \exp Nn\left[-\frac{\hat{q}}{2} + \int Du \log\left(2\cosh\sqrt{\hat{q}}u\right)\right]$$

Second, we simplify the terms that are raised to the power αN:

$$\int \prod_a \frac{\mathrm{d}x^a\mathrm{d}\hat{x}^a}{2\pi} \prod_a \exp\left(-\beta\, f\left(x^a\right)\right) \prod_a \exp\left(i\hat{x}^a x^a\right) \exp\left(-\frac{1}{2}\sum_{ab}\hat{x}^a\hat{x}^b q\right) =$$

$$= \int \prod_a \frac{\mathrm{d}x^a\mathrm{d}\hat{x}^a}{2\pi} \prod_a \exp\left(-\beta\, f\left(x^a\right)\right) \prod_a \exp\left(i\hat{x}^a x^a\right) \exp\left(-\frac{1}{2}\left(1-q\right)\sum_a\left(\hat{x}^a\right)^2 - \frac{q}{2}\left(\sum_a \hat{x}^a\right)^2\right)$$

$$= \int Du \left(\int \frac{\mathrm{d}x\mathrm{d}\hat{x}}{2\pi} e^{-\beta\, f(x)} e^{i\hat{x}x} \exp\left(-\frac{1}{2}\left(1-q\right)\hat{x}^2 + iu\sqrt{q}\hat{x}\right)\right)^n$$

$$= \int Du \left(\int \frac{\mathrm{d}x}{\sqrt{2\pi}} e^{-\beta\, f(x)} \frac{\exp\left(-\frac{(x+u\sqrt{q})^2}{2(1-q)}\right)}{\sqrt{1-q}}\right)^n$$

Finally we can write the saddle point expression for the replicated partition function:

$$\langle Z^n\rangle \simeq \exp\left[Nn\left(\frac{q\hat{q}}{2} - \frac{\hat{q}}{2} + \int Du \log\left(2\cosh\sqrt{\hat{q}}u\right) + \alpha\int Du \log\int \exp\left(-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)\right)\right)\right] = \exp\left(NnG\right)$$

where it is useful to write the action G as the sum of three terms

$$G = \frac{\hat{q}}{2}(q-1) + G_S + \alpha G_E$$

The entropic contribution G_S reads

$$G_S = \int Du \log\left(2\cosh\sqrt{\hat{q}}u\right)$$

and the energetic one G_E

$$G_E = \int Du \log \int \exp\left(-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)\right)$$

The replicated partition function can then be computed in the limit $N \to \infty$ by solving the saddle point equations $\frac{\partial G}{\partial \hat{q}} = 0$ and $\frac{\partial G}{\partial q} = 0$. The derivatives of G_S and G_E read

$$\frac{\partial G_S}{\partial \hat{q}} = \int Du \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}\left[-\frac{\beta}{2}f'\left(x\sqrt{1-q}+u\sqrt{q}\right)\left(\frac{u}{\sqrt{q}}-\frac{x}{\sqrt{1-q}}\right)\right]}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}}$$

and

$$\frac{\partial G_E}{\partial q} = \int Du \frac{u}{2\sqrt{\hat{q}}}\tanh\left(u\sqrt{\hat{q}}\right) = \frac{1}{2}\left(1 - \int Du \tanh^2\left(u\sqrt{\hat{q}}\right)\right)$$

Setting to zero these derivatives we get the saddle point equations for q and $\hat{q}$

$$\begin{aligned}
q &= \int Du \tanh^2\left(u\sqrt{\hat{q}}\right) \\
\hat{q} &= -\alpha\beta \int Du \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}\left[-\frac{\beta}{2}f'\left(x\sqrt{1-q}+u\sqrt{q}\right)\left(\frac{u}{\sqrt{q}}-\frac{x}{\sqrt{1-q}}\right)\right]}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}} = \\
&= -\frac{\alpha}{\sqrt{1-q}} \int Du \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}\left[\frac{1}{\sqrt{1-q}}+x\left(\frac{u}{\sqrt{q}}-\frac{x}{\sqrt{1-q}}\right)\right]}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}}
\end{aligned}$$

In the limit of large β we need to rescale the order parameters to obtain finite quantities. By setting $q = 1 - \frac{\delta q}{\beta}$, we find for the last equation

$$\hat{q} = \frac{\alpha\beta^2}{\delta q}\int Du\ \left[\text{argmax}_x\left(-\frac{x^2}{2} - \log\left(1+\exp\left(-2\gamma\left(x\sqrt{\delta q}+u\right)\right)\right)\right)\right]^2$$

Once the saddle point equations are solved numerically, we can compute the minimum energy (minimum loss) and the entropy at low temperature. We have:

$$E = -\frac{\partial G}{\partial \beta} = -\alpha\frac{\partial G_E}{\partial \beta} = \alpha\int Du \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)} f\left(x\sqrt{1-q}+u\sqrt{q}\right)}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+u\sqrt{q}\right)}}$$

In the limit of large β, with $q = 1 - \frac{\delta q}{\beta}$, we find

$$E = \alpha\int Du\, f\left(x^*\left(u\right)\right)$$

where

$$x^{*}\left(u\right)\equiv\text{argmax}_{x}\left(-\frac{x^{2}}{2}-\log\left(1+\exp\left(-2\gamma\left(x\sqrt{\delta q}+u\right)\right)\right)\right)$$

We can compute the entropy using the relation $S=G+\beta E$.

In figure 2 of the main text we show the behavior of the energy vs the loading α. As one may observe, up to relatively large values of α the energy is extremely small, virtually equal zero for any accessible size N.

Having established that by minimizing the cross-entropy one ends up in regions of perfect classification where the error loss function is zero, it remains to be understood which type of configurations of weights are found. Does the CE converge to a typical zero energy configuration of the error function (i.e. an isolated point-like solution in the weight space) or does it converge to the rare regions of high local entropy?

The answer to this question is that the CE does in fact focus on the HLE subspaces.

B. The local entropy around CE ground states

In order to show this analytically we need to be able to count how many zero error configurations exist in a region close to a typical minima of the CE loss. Following ref. [4], this computation can be done by averaging with the CE Gibbs measure the entropy of the Number of Errors loss function.

Let's call f and g the cross-entropy and error loss functions per pattern, respectively. We need to evaluate the most probable value of

$$\log Z_{FP}=\log\frac{\int\prod_{i}\mathrm{d}\mu\left(w_{i}\right)\prod_{\mu}\exp\left(-\beta f\left(\sum_{i}\frac{w_{i}\xi_{i}^{\mu}}{\sqrt{N}}\right)\right)\log\left[\int\prod_{i}\mathrm{d}\mu\left(v_{i}\right)\prod_{\mu}\exp\left(-\beta' g\left(\sum_{i}\frac{v_{i}\xi_{i}^{\mu}}{\sqrt{N}}\right)\right)\delta\left(pN-\sum_{i}w_{i}v_{i}\right)\right]}{\int\prod_{i}\mathrm{d}\mu\left(w_{i}\right)\prod_{\mu}\exp\left(-\beta f\left(\sum_{i}\frac{w_{i}\xi_{i}^{\mu}}{\sqrt{N}}\right)\right)}$$

which can be computed by replica approach as we have done for Z, yielding the local entropy $\mathcal{E}_{D}=\langle\log Z_{FP}\rangle$. In the above expression pN is the constrained overlap between the different minima, which is trivially related to the Hamming distance DN by $D=\frac{1}{2}\left(1-p\right)$. We will need to perform twice the replica trick, one to extract the most probable value of $\log Z_{FP}$ (index n) and one to linearize the log inside the integral (index r),

$$\begin{aligned}&\langle Z_{FP}^{n,r}\rangle=\\&\left\langle\int\prod_{ai}\mathrm{d}\mu\left(w_{i}^{a}\right)\prod_{\mu a}\exp\left(-\beta f\left(\sum_{i}\frac{w_{i}^{a}\xi_{i}^{\mu}}{\sqrt{N}}\right)\right)\times\right.\\&\left.\times\frac{1}{r}\left(\int\prod_{ci}\mathrm{d}\mu\left(v_{i}\right)\prod_{\mu c}\exp\left(-\beta' g\left(\sum_{i}\frac{v_{i}^{c}\xi_{i}^{\mu}}{\sqrt{N}}\right)\right)\delta\left(pN-\sum_{i}w_{i}^{a=1}v_{i}^{c}\right)-1\right)\right\rangle_{\xi}\end{aligned}$$

This quantity is computed for n,r integer ($a,b=1\ldots,n$ and $c,d,=1,\ldots,r$) and eventually the analytic continuation $n,r\to0$ is taken. We will do this under the replica symmetric (RS) assumption which for small distances D is expected to by exact. For the sake of completeness, we report hereafter all the main steps of the calculation.

We need to compute

$$\begin{aligned}\langle Z_{FP}^{n,r}\rangle=&\int\prod_{ai}\mathrm{d}\mu\left(w_{i}^{a}\right)\int\prod_{a\mu}\frac{\mathrm{d}x^{a\mu}\mathrm{d}\hat{x}^{a\mu}}{2\pi}\prod_{\mu a}\exp\left(-\beta f\left(x^{a\mu}\right)\right)\prod_{a,\mu}e^{i\hat{x}^{a\mu}x^{a\mu}}\times\\&\times\int\prod_{ci}\mathrm{d}\mu\left(v_{i}^{c}\right)\int\prod_{c\mu}\frac{\mathrm{d}y^{c\mu}\mathrm{d}\hat{y}^{c\mu}}{2\pi}\prod\exp\left(-\beta' g\left(y^{c\mu}\right)\right)\times\\&\times\prod_{\mu}\left\langle e^{i\hat{x}^{a\mu}\sum_{ia}\frac{w_{i}^{a}\xi_{i}^{\mu}}{\sqrt{N}}+i\hat{y}^{c\mu}\sum_{ic}\frac{v_{i}^{c}\xi_{i}^{\mu}}{\sqrt{N}}}\right\rangle_{\xi}\prod_{c}\delta\left(pN-\sum_{i}w_{i}^{a=1}v_{i}^{c}\right)\end{aligned}$$

The average over the patterns is factorized and can be easily performed. Upon expanding the results for large N, the term in the brackets reads

$$\exp\left(-\frac{1}{2N}\sum_i\left(\sum_a w_i^a\hat{x}^{a\mu}+\sum_c v_i^c\hat{y}^{c\mu}\right)^2\right)=$$
$$=\exp\left(-\frac{1}{2}\sum_{ab}\hat{x}^{a\mu}\hat{x}^{b\mu}\frac{1}{N}\sum_i w_i^a w_i^b-\frac{1}{2}\sum_{cd}\hat{x}^{c\mu}\hat{x}^{d\mu}\frac{1}{N}\sum_i v_i^c v_i^d-\sum_{ac}\hat{x}^{a\mu}\hat{y}^{c\mu}\frac{1}{N}\sum_i w_i^a v_i^c\right)$$

Introducing the order parameters corresponding to the different overlaps we find for the total expression

$$\begin{aligned}\langle Z_{FP}^{n,r}\rangle=&\int\prod_{a>b}\frac{\mathrm{d}q^{ab}\mathrm{d}\hat{q}^{ab}N}{2\pi}\int\prod_{c>d}\frac{\mathrm{d}s^{cd}\mathrm{d}\hat{s}^{cd}N}{2\pi}\prod_{a>b}e^{-N\hat{q}^{ab}q^{ab}}\prod_{c>d}e^{-N\hat{s}^{cd}s^{cd}}\int\prod_c\frac{\mathrm{d}\hat{p}^cN}{2\pi}\prod_c e^{-N\hat{p}^cp}\times\\
&\times\int\prod_{a>1,c}\frac{\mathrm{d}t^{ac}\mathrm{d}\hat{t}^{ac}}{2\pi}N\prod_{a>1,c}e^{-N\hat{t}^{ac}t^{ac}}\times\\
&\times\int\prod_{ai}\mathrm{d}\mu\left(w_i^a\right)\int\prod_{ci}\mathrm{d}\mu\left(v_i^c\right)\prod_{a>b}e^{\hat{q}^{ab}\sum_i w_i^a w_i^b}\prod_{c>d}e^{\hat{s}^{cd}\sum_i v_i^c v_i^d}\prod_c e^{\hat{p}^c\sum_i w_i^{a=1}v_i^c}\prod_{a>1,c}e^{\hat{t}^{ac}\sum_i w_i^a v_i^c}\times\\
&\times\left(\int\prod_a\frac{\mathrm{d}x^a\mathrm{d}\hat{x}^a}{2\pi}\int\prod_c\frac{\mathrm{d}y^c\mathrm{d}\hat{y}^c}{2\pi}\prod_a e^{-\beta f(x^a)}\prod_c e^{-\beta' g(x^c)}\prod_a e^{ix^a\hat{x}^a}\times\right.\\
&\left.\times\prod_c e^{iy^c\hat{y}^c}e^{-\frac{1}{2}\sum_{ab}\hat{x}^a\hat{x}^b q^{ab}-\frac{1}{2}\sum_{cd}\hat{y}^c\hat{y}^d s^{cd}-\sum_c\hat{x}^1\hat{y}^c p-\sum_{a>1,c}\hat{x}^a\hat{y}^c t^{ac}}\right)^{\alpha N}\end{aligned}$$

In order to proceed, we search the solutions of the saddle point equations in the RS subspace, $q^{ab}=q$, $\hat{q}^{ab}=\hat{q}$, $s^{ab}=s$, etc. The various factors can be simplified as follows:

$$\begin{aligned}\prod_i e^{\sum_{a>b}\hat{q}^{ab}w_i^a w_i^b}&=\prod_i e^{\frac{\hat{q}}{2}\left[\left(\sum_a w_i^a\right)^2-\sum_a\left(w_i^a\right)^2\right]}=e^{-\frac{Nn\hat{q}}{2}}\prod_i e^{\frac{\hat{q}}{2}\left(\sum_a w_i^a\right)^2},\\
\prod_i e^{\sum_{c>d}\hat{s}^{ab}v_i^c v_i^d}&=e^{-\frac{Nr\hat{s}}{2}}\prod_i e^{\frac{\hat{s}}{2}\left(\sum_c v_i^a\right)^2},\\
\prod_i e^{\sum_c\hat{p}^c w_i^{a=1}v_i^c}&=\prod_i e^{\hat{p}\sum_c w_i^{a=1}v_i^c},\\
\prod_i e^{\sum_{a>1,c}\hat{t}^{ac}w_i^a v_i^d}&=\prod_i e^{\hat{t}\left(\sum_{a>1}w_i^a\right)\left(\sum_c v_i^c\right)}=\prod_i e^{\hat{t}\left(\sum_a w_i^a\right)\left(\sum_c v_i^c\right)-\hat{t}w_i^{a=1}\left(\sum_c v_i^c\right)},\end{aligned}$$

and

$$\begin{aligned}-\sum_{a>b}\hat{q}^{ab}q^{ab}&=-\frac{n\left(n-1\right)}{2}\hat{q}q\simeq\frac{n}{2}\hat{q}q,\\
-\sum_{c>d}\hat{s}^{cd}s^{cd}&=-\frac{r\left(r-1\right)}{2}\hat{s}s\simeq\frac{r}{2}\hat{s}s,\\
-\sum_c\hat{p}^c p&=-r\hat{p}p,\\
-\sum_{a>1,c}\hat{t}^{ac}t^{ac}&=-\left(n-1\right)r\hat{t}t\simeq r\hat{t}t.\end{aligned}$$

A series of further simplifications are needed in order to write the $\langle Z_{FP}^{n,r}\rangle$ in the appropriate saddle point form. The terms containing the integrals over w and v become factorized

$$\left(\int\prod_a \mathrm{d}\mu\left(w^a\right)\int\prod_c \mathrm{d}\mu\left(v^c\right)\, e^{-\frac{n}{2}\hat{q}-\frac{r}{2}\hat{s}}\int Dz\, e^{z\sqrt{\hat{q}-\hat{t}}\sum_a w^a}\int Du\, e^{u\sqrt{\hat{s}-\hat{t}}\sum_c v^c}\int Dx\, e^{x\sqrt{\hat{t}}\left(\sum_a w^a+\sum_c v^c\right)}\times\right.$$
$$\left.\times e^{\left(\hat{p}-\hat{t}\right)w^{a=1}\sum_c v^c}\right)^N=$$
$$=\left(\int DzDuDx\left(\int \mathrm{d}\mu\left(w\right)e^{-\frac{\hat{q}}{2}+w\left(z\sqrt{\hat{q}-\hat{t}}+x\sqrt{\hat{t}}\right)}\right)^{n-1}\times\right.$$
$$\left.\times\int \mathrm{d}\mu\left(w^{a=1}\right)e^{-\frac{\hat{q}}{2}+w^{a=1}\left(z\sqrt{\hat{q}-\hat{t}}+x\sqrt{\hat{t}}\right)}\left(\int \mathrm{d}\mu\left(v\right)\, e^{\left(u\sqrt{\hat{s}-\hat{t}}+x\sqrt{\hat{t}}\right)v+\left(\hat{p}-\hat{t}\right)w^{a=1}v}\right)^r\right)^N$$

where we have kept the notation $w^{a=1}$ just for the sake of clarity. Being an integration variable we now drop it. By summing over w and v and with some straightforward change of variables we get

$$\int DzDuDx\frac{\sum_{w=\pm1}e^{w\left(z\sqrt{\hat{q}-\hat{t}}+x\sqrt{\hat{t}}\right)}\left(-\frac{\hat{s}}{2}+\log 2\cosh\left(\left(u\sqrt{\hat{s}-\hat{t}}+x\sqrt{\hat{t}}\right)v+\left(\hat{p}-\hat{t}\right)w\right)\right)}{\sum_{w=\pm1}e^{w\left(z\sqrt{\hat{q}-\hat{t}}+x\sqrt{\hat{t}}\right)}}$$
$$=-\frac{\hat{s}}{2}+\int Dz\frac{\sum_{w=\pm1}e^{wz\sqrt{\hat{q}}}\left(\int DuDx\log\cosh\left(u\sqrt{\hat{s}-\hat{t}}+\sqrt{\hat{t}}\left(\sqrt{\frac{\hat{t}}{\hat{q}}}z+\sqrt{\frac{\hat{q}-\hat{t}}{\hat{q}}}x\right)+\left(\hat{p}-\hat{t}\right)w\right)\right)}{\sum_{w=\pm1}e^{wz\sqrt{\hat{q}}}}=$$
$$=-\frac{\hat{s}}{2}+\int Dz\frac{\sum_{w=\pm1}e^{wz\sqrt{\hat{q}}}\int D\phi\log\cosh\left(\phi\sqrt{\frac{(\hat{s}-\hat{t})\hat{q}+(\hat{q}-\hat{t})\hat{t}}{\hat{q}}}+\frac{\hat{t}}{\sqrt{\hat{q}}}z+\left(\hat{p}-\hat{t}\right)w\right)}{2\cosh z\sqrt{\hat{q}}}=$$

For the integral containing the dependence on f and g we find similar simplifications.

$$\int DzDh\left(\int\frac{\mathrm{d}x\mathrm{d}\hat{x}}{2\pi}e^{-\beta f(x)-\frac{1-q}{2}\hat{x}^2+i\hat{x}\left(z\sqrt{q-t}+h\sqrt{t}+x\right)}\right)^{n-1}\times$$
$$\times\left(\int\prod_c\frac{\mathrm{d}y^c\mathrm{d}\hat{y}^c}{2\pi}\int\frac{dxd\hat{x}}{2\pi}e^{-\beta f(x)-\frac{1-q}{2}\hat{x}^2+i\hat{x}\left(z\sqrt{q-t}+h\sqrt{t}+x+i(p-t)\sum_c\hat{y}^c\right)}\times\right.$$
$$\left.\times\prod_c e^{-\beta' g(y^c)-\frac{1-s}{2}\sum_c(\hat{y}^c)^2-\frac{s-t}{2}\left(\sum_c\hat{y}^c\right)^2+i\sum_c\hat{y}^c\left(h\sqrt{t}+y^c\right)}\right)=$$
$$=\int DzDh\frac{1}{Z_0}\int\prod_c\frac{\mathrm{d}y^c\mathrm{d}\hat{y}^c}{2\pi}\int\frac{\mathrm{d}x}{\sqrt{2\pi}}\frac{e^{-\beta f(x)}}{\sqrt{1-q}}e^{-\frac{1}{2(1-q)}\left[\left(z\sqrt{q-t}+h\sqrt{t}+x\right)^2-(p-t)^2\left(\sum_c\hat{y}^c\right)^2+2i\left(z\sqrt{q-t}+h\sqrt{t}+x\right)(p-t)\sum_c\hat{y}^c\right]}\times$$
$$\times\prod_c e^{-\beta' g(y^c)-\frac{1-s}{2}\sum_c(\hat{y}^c)^2-\frac{s-t}{2}\left(\sum_c\hat{y}^c\right)^2+i\sum_c\hat{y}^c\left(h\sqrt{t}+y^c\right)}=\dots$$

where we have used the notation

$$Z_0\equiv\int\frac{\mathrm{d}x}{\sqrt{2\pi}}\frac{e^{-\beta f(x)}}{\sqrt{1-q}}e^{-\frac{1}{2(1-q)}\left(z\sqrt{q-t}+h\sqrt{t}+x\right)^2}$$

Continuing the computation, we can linearize the terms in $\hat{y}^c$ with an auxiliary integral, factor the terms with the

index c, take the $r \to 0$ limit and perform explicitly two integrals:

$$\begin{aligned}
\cdots &= \int DzDh \frac{1}{Z_0} \int Du \int \frac{\mathrm{d}x}{\sqrt{2\pi}} \frac{e^{-\beta f(x)}}{\sqrt{1-q}} e^{-\frac{1}{2(1-q)}\left(z\sqrt{q-t}+h\sqrt{t}+x\right)^2} \times \\
&\times \left(\int \frac{\mathrm{d}y \mathrm{d}\hat{y}}{2\pi} e^{-\beta' g(y) - \frac{1-s}{2}\hat{y}^2 + i\hat{y}\left(\sqrt{s-t-\frac{(p-t)^2}{1-q}}u + h\sqrt{t} + y - \frac{\left(z\sqrt{q-t}+h\sqrt{t}+x\right)(p-t)}{1-q}\right)} \right)^r = \\
&= \int DzDh \frac{1}{Z_0} \int Du \int \frac{dx}{\sqrt{2\pi}} \frac{e^{-\beta f(x)}}{\sqrt{1-q}} e^{-\frac{1}{2(1-q)}\left(z\sqrt{q}+x\right)^2} \times \\
&\times \log \int \frac{\mathrm{d}y}{\sqrt{2\pi}} \frac{e^{-\beta' g(y)}}{\sqrt{1-s}} e^{-\frac{1}{\sqrt{2(1-s)}}\left[u\sqrt{s-t-\frac{(p-t)^2}{1-q}} + \sqrt{t}\left(z\sqrt{\frac{t}{q}} + h\sqrt{\frac{q-t}{q}}\right) + y - \frac{z\sqrt{q}(p-t)}{1-q}\right]^2} = \\
&= \int Dz \frac{1}{Z_0'} \int \frac{\mathrm{d}x}{\sqrt{2\pi}} \frac{e^{-\beta f(x)}}{\sqrt{1-q}} e^{-\frac{1}{2(1-q)}\left(z\sqrt{q}+x\right)^2} \times \\
&\times \int Du \log \int \frac{\mathrm{d}y}{\sqrt{2\pi}} \frac{e^{-\beta' g(y)}}{\sqrt{1-s}} e^{-\frac{1}{\sqrt{2(1-s)}}\left[u\sqrt{s-t-\frac{(p-t)^2}{1-q}-\frac{t(q-t)}{q}} + \frac{t}{\sqrt{q}}z - \frac{z\sqrt{q}(p-t)}{1-q} + y\right]^2}
\end{aligned}$$

where

$$Z_0' \equiv \int \frac{\mathrm{d}x}{\sqrt{2\pi}} \frac{e^{-\beta f(x)}}{\sqrt{1-q}} e^{-\frac{1}{2(1-q)}\left(z\sqrt{q}+x\right)^2}$$

The local entropy $\mathcal{E}_D = \langle \log Z_{FP} \rangle$ is nothing but the total exponent for the saddle point equations, which after some additional changes of variables can eventually be written as

$$\mathcal{E}_D = \frac{1}{2}\hat{s}s - \hat{p}p + t\hat{t} - \frac{\hat{s}}{2} + G_S + \alpha G_E$$

where

$$G_S = \int Dz \frac{\sum_{w=\pm 1} e^{wz\sqrt{\hat{q}}} \int D\phi \log\left[2\cosh\left(\left(\sqrt{\hat{s} - \frac{\hat{t}^2}{\hat{q}}}\right)\phi + \frac{\hat{t}}{\sqrt{\hat{q}}}z + \left(\hat{p} - \hat{t}\right)w\right)\right]}{2\cosh\left(z\sqrt{\hat{q}}\right)}$$

and

$$G_E = \int Dz \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)} \int Du \log[\int Dy\, e^{-\beta' g\left(y\sqrt{1-s}+\frac{t}{\sqrt{q}}z+\frac{p-t}{\sqrt{1-q}}x+u\sqrt{\omega}\right)}]}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)}}$$

where we have defined $\omega = s - t - \frac{(p-t)^2}{1-q} + \frac{t(q-t)}{q}$.

If we now take the limit $\beta' \to \infty$ and plug in the expression for the error loss function $g(x) = \Theta(-x)$, we can eliminate one integral

$$\begin{aligned}
&\lim_{\beta' \to \infty} \int Dy\, e^{-\beta' \Theta\left(-\left(y\sqrt{1-s}+\frac{t}{\sqrt{q}}z+\frac{p-t}{\sqrt{1-q}}x+u\sqrt{\omega}\right)\right)} = \\
&= \int Dy\, \Theta\left(y\sqrt{1-s} + \frac{t}{\sqrt{q}}z + \frac{p-t}{\sqrt{1-q}}x + w\sqrt{\omega}\right) = H\left(-\frac{\frac{t}{\sqrt{q}}z + \frac{p-t}{\sqrt{1-q}}x + u\sqrt{\omega}}{\sqrt{1-s}}\right)
\end{aligned}$$

and the expression of G_E simplifies to

$$G_E = \int Dz \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)} \int Du \log H\left(-\frac{\frac{t}{\sqrt{q}}z+\frac{p-t}{\sqrt{1-q}}x+u\sqrt{\omega}}{\sqrt{1-s}}\right)}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)}}$$

In order to compute the entropy for a given distance $D = (1-p)/2$, we need to solve the saddle point equations with respect to $s, \hat{s}, t, \hat{t}, \hat{p}$ with the values of q and $\hat{q}$ obtained by solving the equations for the CE loss function. The saddle point equations can be written as

$$
\begin{aligned}
0 &= \frac{\partial \mathcal{E}_D}{\partial \hat{s}} = \frac{1}{2}(s-1) + \frac{\partial G_S}{\partial \hat{s}} \\
0 &= \frac{\partial \mathcal{E}_D}{\partial \hat{p}} = -p + \frac{\partial G_S}{\partial \hat{p}} \\
0 &= \frac{\partial \mathcal{E}_D}{\partial \hat{t}} = t + \frac{\partial G_S}{\partial \hat{t}} \\
0 &= \frac{\partial \mathcal{E}_D}{\partial s} = \frac{\hat{s}}{2} + \alpha \frac{\partial G_E}{\partial s} \\
0 &= \frac{\partial \mathcal{E}_D}{\partial t} = t + \alpha \frac{\partial G_E}{\partial t}
\end{aligned}
$$

with

$$
\begin{aligned}
\frac{\partial G_S}{\partial \hat{s}} &= \int Dz \frac{\sum_{w=\pm 1} e^{wz\sqrt{\hat{q}}} \int D\phi \tanh\left(\phi\sqrt{\hat{s}-\frac{\hat{t}^2}{\hat{q}}} + z\frac{\hat{t}}{\sqrt{\hat{q}}} + (\hat{p}-\hat{t})\,w\right) \frac{\phi}{2\sqrt{\hat{s}-\frac{\hat{t}^2}{\hat{q}}}}}{2\cosh z\sqrt{\hat{q}}}, \\
\frac{\partial G_S}{\partial \hat{p}} &= \int Dz \frac{\sum_{w=\pm 1} e^{wz\sqrt{\hat{q}}} \int D\phi \tanh\left(\phi\sqrt{\hat{s}-\frac{\hat{t}^2}{\hat{q}}} + z\frac{\hat{t}}{\sqrt{\hat{q}}} + (\hat{p}-\hat{t})\,w\right) w}{2\cosh z\sqrt{\hat{q}}}, \\
\frac{\partial G_S}{\partial \hat{t}} &= \int Dz \frac{\sum_{w=\pm 1} e^{wz\sqrt{\hat{q}}} \int D\phi \tanh\left(\phi\sqrt{\hat{s}-\frac{\hat{t}^2}{\hat{q}}} + z\frac{\hat{t}}{\sqrt{\hat{q}}} + (\hat{p}-\hat{t})\,w\right)}{2\cosh z\sqrt{\hat{q}}} \left[-\frac{\hat{t}\phi}{\hat{q}\sqrt{\hat{s}-\frac{\hat{t}^2}{\hat{q}}}} + z\frac{\hat{t}}{\sqrt{\hat{q}}} - w\right], \\
\frac{\partial G_E}{\partial s} &= \int Dz \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)} \int Du\, \mathcal{G}\left(\frac{\frac{t}{\sqrt{q}}z + \frac{p-t}{\sqrt{1-q}}x + u\sqrt{\omega}}{\sqrt{1-s}}\right)\left[\frac{u}{2\sqrt{1-s}\sqrt{\omega}} + \frac{\frac{t}{\sqrt{q}}+x\frac{p-t}{\sqrt{1-q}}+u\sqrt{\omega}}{2(1-s)^{3/2}}\right]}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)}}, \\
\frac{\partial G_E}{\partial t} &= \int Dz \frac{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)} \int Du\, \mathcal{G}\left(\frac{\frac{t}{\sqrt{q}}z + \frac{p-t}{\sqrt{1-q}}x + u\sqrt{\omega}}{\sqrt{1-s}}\right)\left[-\frac{x}{\sqrt{1-q}} + \frac{z}{\sqrt{q}} + u\frac{1}{\sqrt{\omega}}\frac{pq-t}{q(1-q)}\right]\frac{1}{\sqrt{1-s}}}{\int Dx\, e^{-\beta f\left(x\sqrt{1-q}+z\sqrt{q}\right)}}
\end{aligned}
$$

where we have defined $\mathcal{G}(-x) \equiv \frac{\partial}{\partial x}\log H[-x]$.

The results are reported in fig. 3 of the main text. We may observe that the minima of the CE are indeed surrounded by an exponential number of zero error solutions. In other words, the CE focuses on HLE regions.

It is clear from the figure that γ needs to be sufficiently large for this phenomenon to occur. On the other hand, in the limit $\gamma \to \infty$ the space of solutions is again dominated by the isolated ones; we thus expect the existence of an optimal value of γ, depending on the parameters α and D. We set $\alpha = 0.4$ and used two values of D, 0.005 and 0.02, and measured the normalized local entropy $\mathcal{E}_D - \mathcal{E}_D^{\max}$ where $\mathcal{E}_D^{\max} = -D\log D - (1-D)\log(1-D)$ is the upper bound corresponding to the case $\alpha = 0$ (gray curve in fig. 3 of the main text). The results are shown in fig. S1. Numerical issues prevented us from reaching the optimal γ at $D = 0.005$ (the main plot, in log scale, shows that the curve is still growing), and the left inset (same data as the main plot, but not in log scale) shows that there is a whole region where the local entropy is extremely close to optimal (what we would call a dense region). At a larger distance, $D = 0.02$, we could find the optimum (denoted with a dot) at a lower local entropy (consistently with fig. 3 of the main text), but it is also clear that the "good" region of γ is rather large since the curve is very flat. In the limit of $\gamma \to \infty$ the curves would tend to -0.054 and -0.028 for the $D = 0.02$ and $D = 0.005$ cases, respectively.

The physical significance of these "optimal γ regions" and their relation with the local geometry of the landscape is not obvious. The main reason for this is that a fine description of the local geometry of the HLE regions is still an open problem – circumstantial evidence from theoretical considerations and numerical simulations suggests that they are rather complex structures [1], possibly with a multifractal nature. At the present time, this is pure speculation. On the practical side, however, a local search algorithm that explores the landscape at low γ and gradually increases it should eventually reach this "good γ" plateau that leads it toward a HLE region; further increasing γ would then be unlikely to drive it out, unless very strong drift terms are present [5].

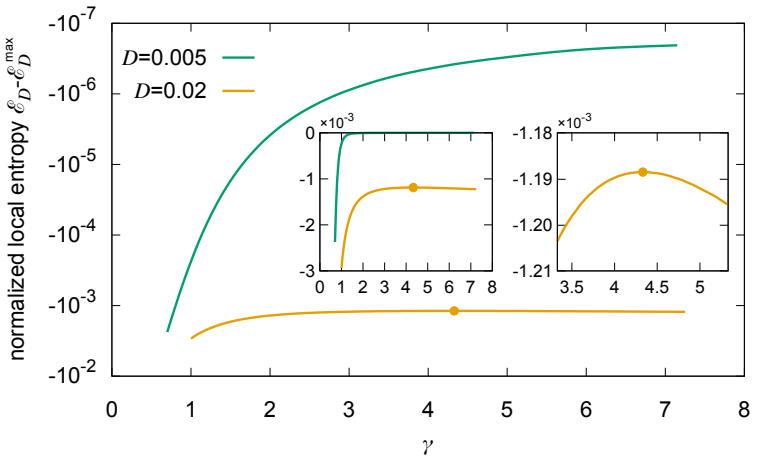

Figure S1. Normalized local entropy as a function of γ, for a binary perceptron trained with the CE loss at $\alpha = 0.4$ and two values of the distance D. The main plot shows the results in logarithmic scale, the left inset is the same plot in linear scale. For $D = 0.005$ we could not find the optimal γ due to numerical issues; for $D = 0.02$ the optimum is shown with a circle, although the curve looks rather flat in an extended region: the right inset shows an enlargement that makes the appearance of a maximum evident.

C. RS stability and zero entropy condition

In order to corroborate the validity of the RS solution, we need to check two necessary conditions: the entropy of the CE model is positive and the replica symmetric solution is stable. In other words we can focus on values of the parameters α, γ, β such that both conditions are met.

Following ref. [3], the stability is verified if the following condition is met

$$\alpha \gamma_E \gamma_S < 1$$

where γ_E and γ_S are the two eigenvalues of the Hessian matrix computed at the RS saddle point. We find

$$\gamma_S = \int Dz \left[1 - \tanh(z\sqrt{\hat{q}})\right]^2$$

and

$$\gamma_E = \frac{1}{(1-q)^2} \int Dz \left[1 - \left(\langle x^2 \rangle_z - \langle x \rangle_z^2\right)\right]^2$$

where the averages can be expressed through the quantity

$$\overline{\tilde{x}^k(z)} = \frac{\int \frac{dx d\tilde{x}}{2\pi} \tilde{x}^k e^{-\frac{1}{2}(1-q)\tilde{x}^2 + i\tilde{x}\left(z\sqrt{q}+x\right)} e^{-\beta f(x)}}{\int \frac{dx d\tilde{x}}{2\pi} e^{-\frac{1}{2}(1-q)\tilde{x}^2 + i\tilde{x}\left(z\sqrt{q}+x\right)} e^{-\beta f(x)}}$$

One finds:

$$\overline{\tilde{x}(z)} = \frac{i}{1-q} \frac{\int \frac{dx}{\sqrt{2\pi}} e^{-\frac{(z\sqrt{q}+x)^2}{2(1-q)}} \left(z\sqrt{q}+x\right) e^{-\beta f(x)}}{\int \frac{dx}{\sqrt{2\pi}} e^{-\frac{(z\sqrt{q}+x)^2}{2(1-q)}} e} = \frac{i}{\sqrt{1-q}} \frac{\int Dx\, x\, e^{-\beta f\left(z\sqrt{q}+x\sqrt{1-q}\right)}}{\int Dx e^{-\beta f\left(z\sqrt{q}+x\sqrt{1-q}\right)}} = \frac{i}{\sqrt{1-q}} \langle x \rangle_z$$

and similarly

$$\overline{\tilde{x}^2(z)} = \frac{1}{(1-q)^2}\frac{\int Dx\ \left(1-q-(1-q)x^2\right)\ e^{-\beta f\left(z\sqrt{q}+x\sqrt{1-q}\right)}}{\int Dx e^{-\beta f\left(z\sqrt{q}+x\sqrt{1-q}\right)}} = \frac{1}{1-q} - \frac{\left\langle x^2\right\rangle_z}{(1-q)}$$

For each α,we can thus identify the values of γ and β for which both the entropy is positive and the solution is stable. In particular, β can be chosen to be quite large, corresponding to energies that are extremely small (RSB is expected to have relatively minor effects at zero temperature).

III. WIDE FLAT MINIMA FOR THE CONTINUOUS CASE

Following ref. [6] and the technique described in sec. I we may analyze the existence of WFM by studying the 1-RSB saddle point equations, with q_1 and y (usually called m in the 1-RSB context) used as control parameters.

The computation of the average of $\langle \log V\rangle_\xi$ over the patterns by the replica method leads to the following saddle point expression in the large N limit

$$\frac{1}{N}\langle \log V\rangle_\xi = \mathrm{extr}_{q_\ell^{ab},\hat{q}_\ell^{ab},E_\ell^a} G\left(\left\{q_\ell^{ab},\ \hat{q}_\ell^{ab},\ E_\ell^a\right\}\right)$$

where

$$G\left(\left\{q_\ell^{ab},\ \hat{q}_\ell^{ab},\ E_\ell^a\right\}\right) = G_S\left(\left\{q_\ell^{ab},\ \hat{q}_\ell^{ab},\ E_\ell^a\right\}\right) + \alpha G_E\left(\left\{q_\ell^{ab}\right\}\right).$$

Given that the distribution of the input patterns is the same for each hidden unit, averages are expected to be independent of ℓ and the dependency on ℓ of the order parameters can be dropped $\left\{q_\ell^{ab},\ \hat{q}_\ell^{ab},\ E_\ell^a\right\} \to \left\{q^{ab},\ \hat{q}^{ab},\ E^a\right\}$. In 1-RSB scheme, once the conjugate order parameters $\left\{\hat{q}^{ab},\ E^a\right\}$ are integrated out, the expressions for G_S and G_E read

$$G_S(q_0,q_1,y) = \frac{1}{2}\left[\frac{1+(y-1)\Delta q_1}{1-q_1+y\Delta q_1} + \ln 2\pi + \left(1-\frac{1}{y}\right)\ln(1-q_1) + \frac{1}{y}\ln(1-q_1+y\Delta q_1)\right]$$
$$G_E(q_0,q_1,y) = \frac{1}{y}\int\prod_{\ell=1}^{K} Dv_\ell \ln\left(\int\prod_{\ell=1}^{K} Du_\ell\left(\Sigma_{(K)}\right)^y\right)$$

where $\Delta q_1 = q_1 - q_0$ and $\Sigma_{(K)}$ is a complicated function of the order parameters which for $K=3$ reads

$$\Sigma_{(3)} = H_1H_2 + H_1H_3 + H_2H_3 - 2H_1H_2H_3$$

with

$$H_\ell = H\left[\sqrt{\frac{\Delta q_1}{1-q_1}}u_\ell + \sqrt{\frac{q_0}{1-q_1}}v_\ell\right].$$

For our WFM computation, q_1 is the constrained overlap between the weight vectors of the m real replicas. q_0 is the only parameter for which we have to solve the saddle point equation. In order to look for the WFM of maximum volume we are interested in the large y limit. In this case the expressions simplify substantially

$$G_S(q_1) = \frac{1}{2}\left[1+\ln 2\pi + \ln(1-q_1)\right]$$
$$G_E(q_0,q_1) = \int\prod_{\ell=1}^{K} Dv_\ell max_{u_1,,u_2,u_3}\left[-\frac{\sum_{\ell=1}^{3}u_\ell^2}{2} + \log\left(\tilde{H}_1\tilde{H}_2 + \tilde{H}_1\tilde{H}_3 + \tilde{H}_2\tilde{H}_3 - 2\tilde{H}_1\tilde{H}_2\tilde{H}_3\right)\right]$$

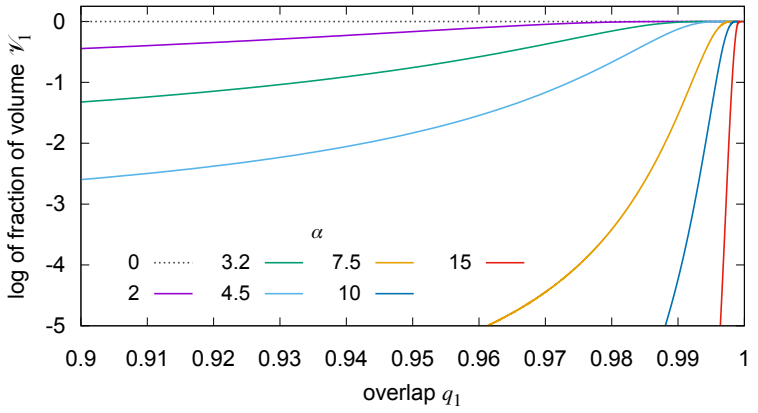

Figure S2. Normalized local entropy $\mathscr{V}_1$ vs q_1, for various values of α, for a tree-like committee machine in the limiting case of a large number of hidden units K. This is the analogous of plot of fig. 1 of the main text, which shows the case of $K = 3$, and the qualitative behavior is indeed very similar.

where $\tilde{H}_\ell \equiv H\left[\sqrt{\frac{d_0}{1-q_1}}u_\ell + \sqrt{\frac{q_0}{1-q_1}}v_\ell\right]$ and $d_0 \equiv y\Delta q_1$. The pre-factor $\frac{1}{y}$ of G_E has been eliminated by a change of variables $u'_\ell = \frac{u_\ell}{\sqrt{y}}$ (with u' then renamed u).

Notice that $G_S(q_1)$ corresponds to the volume at $\alpha = 0$, i.e. to the volume of the weight space just under the spherical constraints, with the real replicas forced to be at an overlap q_1. If WFM exist for positive α, we expect to observe the normalized local entropy $\mathscr{V}_1(q_1) = \alpha G_E(q_1)$ to approach 0 for q_1 sufficiently close to one.

In fig. 1 of the main text we report the values of the WFM volumes vs the overlap q_1, for different values of α. Indeed one may observe that the behavior is qualitatively similar to that of the binary perceptron: WFM exist deep into the RSB region, i.e. besides the RSB states and all the related local minima and saddles, there exist absolute minima that are flat at large distances. We mention that evaluating the max inside the integral is a quite challenging task as the function to be maximized may present multiple maxima. We have tackled this problem by an appropriate sampling technique.

The case $K = 3$ is still relatively close to the perceptron, though the geometrical structure of its minima is already dominated by non convex RSB features for $\alpha > 1.76$. A case that is closer to more realistic NN is $K \gg 1$, which, luckily enough, is easier to study analytically [6].

The 1-RSB expression for G, simplifies to

$$G_S(q_0, q_1, y) = \frac{1}{2}\left[\frac{1 + (y-1)\Delta q_1}{1 - q_1 + y\Delta q_1} + \ln 2\pi + \left(1 - \frac{1}{y}\right)\ln(1-q_1) + \frac{1}{y}\ln(1 - q_1 + y\Delta q_1)\right]$$

$$G_E(q_0, q_1, y) = \frac{1}{y}\int Dv \ln \int Du \left(H\left[\sqrt{\frac{\Delta q_{1\text{eff}}}{1-q_{1\text{eff}}}}u + \sqrt{\frac{q_{0\text{eff}}}{1-q_{1\text{eff}}}}v\right]\right)^y$$

where $q_{0\text{eff}} = 1 - \frac{2}{\pi}\arccos q_0$, $q_{1\text{eff}} = 1 - \frac{2}{\pi}\arccos q_1$ and $\Delta q_{1\text{eff}} = q_{1\text{eff}} - q_{0\text{eff}}$. While the critical capacity diverges with K as $\sqrt{\ln K}$ [7], the value of α at which RSB sets in and the landscape of the minima becomes non trivial remains finite, $\alpha_{RSB} \simeq 2.95$.

As we have done for the $K = 3$, we study the large y limit. We find

$$G_S(q_1) = \frac{1}{2}\left[1 + \ln 2\pi + \ln(1-q_1)\right]$$

$$G_E(q_0, q_1) = \int Dv \max_u \left\{-\frac{u^2}{2} + \log\left[H\left(\sqrt{\frac{\Delta q_1^e}{1-q_1^e}}u + \frac{q_0^e}{1-q_1^e}v\right)\right]\right\}$$

In fig. S2 we observe that WFM are indeed still present.

In order to check the validity of the WFM computation one should check for the stability of the solutions of the saddle point equations by a stability analysis or a 2-RSB computation. For numerical reasons this is a quite difficult

task. We thus decided to follow a different path which is also of algorithmic interest, namely to study the problem by Belief Propagation.

A. Large deviation analysis of the Parity Machine

Here we report the results of the large deviation analysis on the so-called parity machine. Its network structure is identical to the committee machine, except for the output unit which performs the product of the K hidden units instead of taking a majority vote. The outputs of the hidden units are still given by sign activations. Thus, the overall output of the network reads:

$$\sigma_{\text{out}}^{\mu} = \prod_{\ell=1}^{K} \tau_{\ell}.$$

For a given set of patterns, the volume of the weights that correctly classify the patterns is then given by

$$V = \int \prod_{i\ell} \mathrm{d}w_{\ell i} \prod_{\ell} \delta\left(\sum_{i} w_{\ell i}^{2} - \tilde{N}\right) \prod_{\mu} \Theta\left(\sigma^{\mu} \prod_{\ell=1}^{K} \tau_{\ell}^{\mu}\right).$$

The computation proceeds as for the committee machine case, until we find the following expressions for the 1-RSB volume:

$$G_S\left(q_1\right) = \frac{1}{2}\left[\frac{1+\left(y-1\right)\Delta q_1}{1-q_1+y\Delta q_1} + \ln 2\pi + \left(\frac{y-1}{y}\right)\ln\left(1-q_1\right) + \frac{1}{y}\ln\left(1-q_1+y\Delta q_1\right)\right]$$

$$G_E\left(q_1\right) = \frac{1}{y}\int \prod_{\ell=1}^{K} Dv_{\ell} \ln\left[\int \prod_{\ell=1}^{K} Du_{\ell}\left(\sum_{\{\tau_{\ell}\}} \prod_{\ell} H_{\ell}\left(\tau_{\ell}\omega_{\ell}\right)\Theta\left(\prod_{\ell=1}^{K}\tau_{\ell}\right)\right)^{y}\right]$$

where $\Delta q_1 = q_1 - q_0$ and where $\omega_{\ell} = \sqrt{\frac{\Delta q_1}{1-q_1}}u_{\ell} + \sqrt{\frac{q_0}{1-q_1}}v_{\ell}$ and q_0 is fixed by a saddle point equation. The sum over the internal states can be computed for general K, leading to the following final expression for G_E:

$$G_E\left(q_1\right) = \frac{1}{y}\int \prod_{\ell=1}^{K} Dv_{\ell} \ln\left[\int \prod_{\ell=1}^{K} Du_{\ell} \frac{1}{2^{y}}\left(1+\zeta_K\left(\{\omega_{\ell}\}\right)\right)^{y}\right]$$

where $\zeta_K\left(\{\omega_{\ell}\}\right) = (-1)^{K}\prod_{\ell=1}^{K}\left(1-2H_{\ell}\left(\omega_{\ell}\right)\right)$. In the large K limit, $\zeta_K\left(\{\omega_{\ell}\}\right)$ converges rapidly to zero and the expression for G_E simplifies. We can thus compute the volume by optimizing over q_0 for arbitrary y and q_1, and compare it to the volume that one would find for the same distance when no patterns are stored: the log-ratio of the two volumes is constant and equal to $-\alpha\log\left(2\right)$. This shows that the minima never become flat, at any distance scale.

IV. BELIEF PROPAGATION ON A TREE-LIKE COMMITTEE MACHINE WITH CONTINUOUS WEIGHTS: EQUATIONS, LOCAL VOLUME AROUND SOLUTIONS AND ALGORITHMS FOR THE REPLICATED NETWORK

A. BP equations for the committee machine

We can use Belief Propagation (see e.g. refs. [5, 8, 9]) to study the space of the solutions of a tree-like committee machine with continuous weights and random inputs (the outputs can either be random or generated from a rule). The messages in this case are probability density distributions over $\mathbb{R}$, and we will need to ensure normalization by using an additional constraints over the norm of the weights vectors.

The basic factor graph is thus composed of N continuous variable nodes x_i, divided into K groups of N/K variables each; for each pattern, we will have K factor nodes, each one involving one group of N/K variables and an auxiliary binary output variable τ_{ka} (with two indices, one for the hidden unit and one for the pattern), and another factor

node connected to the τ variables and controlling the final output. We will enforce the normalization of the weights by adding an extra field for each variable x_i, as explained below.

As a general notation scheme, we will use the letter h to denote messages from variable nodes to factor nodes, the letter u for messages from factor nodes to variable nodes, and the letter m for non-cavity marginals over variables.

Let's then call $h_{ki\to ka}(x_{ki})$ the cavity message from variable ki (representing a weight with hidden unit index $k\in\{1,\dots,K\}$, weight index $i\in\{1,\dots,N/K\}$, whose value is x_{ki}) to the factor node ka (representing the part of an input pattern a that involves the hidden unit k). The BP equation reads:

$$h_{ki\to ka}(x_{ki}) \propto u_{\mathrm{n}}(x_{ki}) \prod_{b\neq a} u_{kb\to ki}(x_{ki}) \tag{S19}$$

where $u_{kb\to ki}(x_{ki})$ represents a message from another pattern node kb to the variable ki, while $u_{\mathrm{n}}(x_{ki})$ is an external field enforcing the normalization constraint, which is the same for all variables:

$$u_{\mathrm{n}}(x) = e^{-\frac{\chi}{2}x^2} \tag{S20}$$

This method of enforcing normalization is equivalent to using a Dirac delta in the limit of large N, but it's easier to implement in the BP algorithm. In order to set the parameter χ we will need to evaluate the norm of the weights at convergence (detailed below) and adjust χ until the norm matches the desired value, $\sum_{ki} x_{ki}^2 = N$.

The other messages read:

$$u_{ka\to ki}(x_{ki}) = \tag{S21}$$
$$\sum_{\tau_{ka}=\pm 1} h_{ka}(\tau_{ka}) \int \prod_{j\neq i} h_{kj\to ka}(x_{kj})\,\mathrm{d}x_{kj}\;\Theta\left(\tau_{ka}\left(\sum_{j\neq i} x_{kj}\xi_{kj}^{ka} + x_{ki}\xi_{ki}^{ka}\right)\right)$$

where the message h_{ka} goes from the auxiliary output variable τ_{ka} to node ka. Notice that the messages h are assumed to be normalized, but the messages u aren't, because the integral of expression (S21) might diverge.

In the limit of large N, eq. (S21) can be approximated using the central limit theorem, since we're assuming that the input pattern entries are random i.i.d. variables. Therefore, we don't need the full distributions $h_{kj\to ka}$ to perform the integral, only their first and second moments. For simplicity of implementation, though, instead of the moments it is actually more convenient to use the inverse of the variance and the mean rescaled by the variance, which we will denote with ζ and $\tilde{\mu}$, respectively. We will use the following notation:

$$\zeta_{ki\to ka} = \left(\left\langle x_{ki}^2\right\rangle_{h_{ki\to ka}} - \left\langle x_{ki}\right\rangle_{h_{ki\to ka}}^2\right)^{-1} \tag{S22}$$
$$\tilde{\mu}_{ki\to ka} = \zeta_{ki\to ka}\,\left\langle x_{ki}\right\rangle_{h_{ki\to ka}} \tag{S23}$$

With these, we can compute the two auxiliary quantities

$$c_{ka\to ki} = \sum_{j\neq i} \frac{\tilde{\mu}_{kj\to ka}}{\zeta_{kj\to ka}} \xi_{kj}^{ka} \tag{S24}$$
$$v_{ka\to ki} = \sum_{j\neq i} \zeta_{kj\to ka}^{-1} \left(\xi_{kj}^{ka}\right)^2 \tag{S25}$$

which can be simplified by noting that $\left(\xi_{kj}^{ka}\right)^2 = 1$. We can now rewrite eq. (S21) using the central limit theorem:

$$\begin{aligned} u_{ka\to ki}(x_{ki}) &= \sum_{\tau_{ka}=\pm 1} h_{ka}(\tau_{ka}) \int Dz\;\Theta\left(\tau_{ka}\left(c_{ka\to ki} + z\sqrt{v_{ka\to ki}} + x_{ki}\xi_{ki}^{ka}\right)\right) \\ &= \frac{1}{2}\left(1 + h_{ka}\mathrm{erf}\left(\frac{c_{ka\to ki} + x_{ki}\xi_{ki}^{ka}}{\sqrt{2v_{ka\to ki}}}\right)\right) \end{aligned} \tag{S26}$$

where in the last line we have introduced the shorthand notation

$$h_{ka} \equiv h_{ka}(+1) - h_{ka}(-1)$$

because we can represent any message over a binary variable with a single quantity (in this case, the "magnetization" often employed in spin glass literature) and therefore we abuse the notation and identify that message with a single parameter. The context and the presence or absence of the argument will suffice to disambiguate the notation.

In the limit of large N, we observe that $c_{ka\to ki} = O\left(\sqrt{N}\right)$ and $v_{ka\to ki} = O(N)$, and therefore $\frac{c_{ka\to ki}}{\sqrt{v_{ka\to ki}}} = O(1)$, but $\frac{x_{ki}\xi_{ki}^{ka}}{\sqrt{v_{ka\to ki}}} = O\left(\frac{1}{\sqrt{N}}\right)$ since we are assuming $x_{ki} = O(1)$ due to the normalization constraint. We can therefore expand $u_{ka\to ki}$ to the second order, obtaining (up to an irrelevant factor):

$$u_{ka\to ki}(x_{ki}) \propto 1 + U_{ka\to ki}x_{ki} - \frac{1}{2}\left(V_{ka\to ki} - U_{ka\to ki}^2\right)x_{ki}^2 \tag{S27}$$

where

$$U_{ka\to ki} = \frac{\xi_{ki}^{ka}}{\sqrt{v_{ka\to ki}}} \frac{2h_{ka}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\frac{c_{ka\to ki}^2}{v_{ka\to ki}}\right)}{1 + h_{ka}\mathrm{erf}\left(\frac{c_{ka\to ki}}{\sqrt{2v_{ka\to ki}}}\right)} \tag{S28}$$

$$V_{ka\to ki} = \frac{\xi_{ki}^{ka}}{\sqrt{v_{ka\to ki}}} \frac{c_{ka\to ki}}{\sqrt{v_{ka\to ki}}} U_{ka\to ki} + U_{ka\to ki}^2 \tag{S29}$$

Now we can substitute this into eq. (S19):

$$h_{ki\to ka}(x_{ki}) = \frac{1}{z_{ki\to ka}} e^{-\frac{\chi}{2}x_{ki}^2} \prod_{b\neq a}\left(1 + U_{kb\to ki}x_{ki} - \frac{1}{2}\left(V_{ka\to ki} - U_{ka\to ki}^2\right)x_{ki}^2\right)$$

where $z_{ki\to ka}$ is a normalization constant. Expanding this to the second order we finally obtain:

$$h_{ki\to ka}(x_{ki}) = \frac{1}{z_{ki\to ka}} e^{-\frac{1}{2}\left(\chi + \sum_{b\neq a} V_{kb\to ki}\right)x_{ki}^2 + \sum_{b\neq a} U_{kb\to ki}x_{ki}} \tag{S30}$$

and thus we obtain the following simple expression for the parameters of the distribution eqs. (S22)-(S23):

$$\zeta_{ki\to ka} = \chi + \sum_{b\neq a} V_{kb\to ki} \tag{S31}$$

$$\tilde{\mu}_{ki\to ka} = \sum_{b\neq a} U_{kb\to ki} \tag{S32}$$

Also, the normalization constant is:

$$z_{ki\to ka} = \frac{\sqrt{2\pi}}{\sqrt{\lambda_{ki\to ka}}} e^{\frac{1}{2}\frac{\tilde{\mu}_{ki\to ka}^2}{\zeta_{ki\to ka}}} \tag{S33}$$

One can immediately write the corresponding formulas for the non-cavity parameters of the marginal $m_{ki}(x_{ki})$:

$$\zeta_{ki} = \chi + \sum_{b} V_{kb\to ki} \tag{S34}$$

$$\tilde{\mu}_{ki} = \sum_{b} U_{kb\to ki} \tag{S35}$$

$$z_{ki} = \frac{\sqrt{2\pi}}{\sqrt{\lambda_{ki}}} e^{\frac{1}{2}\frac{\tilde{\mu}_{ki}^2}{\zeta_{ki}}} \tag{S36}$$

This form of the equations shows that we can, in fact, parametrize all the distributions as Gaussian, and also that the updates in the equations can be performed efficiently for each node by keeping in memory the non-cavity versions of the parameters ζ and $\tilde{\mu}$ and the parameters U and V, such that updating $\zeta, \tilde{\mu}$ is only a matter of subtracting the previous value of U, V and adding the new one.

We can also note that the computation of U, V in eqs. (S28)-(S29) can be made computationally more efficient by writing $c_{ka\to ki}$ and $v_{ka\to ki}$ in terms of their non-cavity counterparts

$$c_{ka} = \sum_{j} \frac{\tilde{\mu}_{kj\to ka}}{\zeta_{kj\to ka}} \xi_{kj}^{ka} \tag{S37}$$

$$v_{ka} = \sum_{j} \zeta_{kj\to ka}\left(\xi_{kj}^{ka}\right)^2 \tag{S38}$$

and expanding. However, in practice this further approximation does not work well at large values of α when using the Focusing-BP protocol (detailed below, sec. IV C) and at moderate values of N: by avoiding it, we can find solutions at slightly larger values of α (e.g. we can reach $\alpha = 2.7$ instead of 2.6 at $N = 999$, $K = 3$). Relatedly, it is also slightly problematic at large values of the polarization field λ when exploring the space of configurations (see below, sec. IV B). In general terms, the problems arise when the fields become very polarized and the normalization constants become very small. Therefore, in our implementation this approximation is optional.

In order to set the normalization parameter χ, we simply compute the following quantity at the end of each iteration

$$\frac{1}{N}\sum_{ki}\langle x_{ki}^2\rangle = \frac{1}{N}\sum_{ki}\left(\zeta_{ki}^{-1} + \zeta_{ki}^{-2}\tilde{\mu}_{ki}^2\right) \tag{S39}$$

and adjust χ so that this becomes 1. Of course, we also add the criterion that this adjustment needs to be sufficiently small in evaluating whether the BP equations have converged.

The remaining BP equations involve the nodes connecting the auxiliary variables τ_{ka} and enforcing the desired outputs from the committee:

$$u_{ka}\left(\tau_{ka}\right) = \frac{1}{2}\left(1 + \tau_{ka}\mathrm{erf}\left(\frac{c_{ka}}{\sqrt{2v_{ka}}}\right)\right) \tag{S40}$$

$$h_{ka}\left(\tau_{ka}\right) = \sum_{\{\tau_{la}\}_{l\neq k}}\prod_{l\neq k} u_{la}\left(\tau_{la}\right)\Theta\left(\sigma_a\left(\sum_{l\neq k}\tau_{la} + \tau_{ka}\right)\right) \tag{S41}$$

where σ_a is the desired output for pattern a. For small K, we compute h_{ka} exactly (i.e. without using the central limit theorem), which can be done in $O\left(K^2\right)$ time.

Equations (S24), (S25), (S28), (S29), (S31), (S32), (S37), (S38), (S40) and (S41), together with the normalization requirement obtained through eq. (S39), form the full set of BP equations.

The free entropy (also sometimes known as action) of the system can be computed from the usual BP formulas. We get:

$$\phi = \frac{1}{N}\left(\sum_{ki} f_{ki} - \sum_{ka} f_{ka} - \sum_{a} f_a\right) \tag{S42}$$

where:

$$f_{ki} = \log z_{ki} \tag{S43}$$

$$f_{ka} = \sum_{i}\left(\log z_{ki\to ka} - \log z_{ki}\right) \tag{S44}$$

$$f_a = -\log\left(1 + \frac{1}{2}\sigma_a\mathrm{erf}\left(\frac{\sum_k h_{ka}}{\sqrt{2\sum_k\left(1-h_{ka}^2\right)}}\right)\right) \tag{S45}$$

From this, we can compute the entropy by simply accounting for the energetic contribution introduced by the normalization constraint. We also shift it by subtracting the log-volume of the normalized sphere, so that its value is upper bounded by 0

$$S = \phi + \frac{\chi}{2} - \log\left(\sqrt{2\pi e}\right) \tag{S46}$$

B. Exploring the space of solutions around a given configuration

Given a particular configuration $\tilde{W}$, which we assume normalized as $\sum_{ki}\tilde{w}_{ki}^2 = N$, we are interested in exploring the space of solutions at a given distance D from it (for a suitable definition of the distance). Analogously to the norm, in the limit of large N we can control the distance by just adding an extra field to each node (putting it as an extra factor in eq. (S19)), of the form:

$$u_D\left(x_{ki}\right) = e^{-\frac{\lambda}{2}\left(x_{ki}-\tilde{w}_{ki}\right)^2}$$

By varying the auxiliary parameter λ between 0 and ∞ we can restrict ourselves to smaller and smaller regions around $\tilde{w}$. Adding this extra field in practice just amounts at adding two terms $\lambda\tilde{w}_{ki}$ and λ to the expressions of $\tilde{\mu}_{ki\to ka}$ and $\zeta_{ki\to ka}$, respectively (eqs. (S31)-(S32)).

For convenience, we actually abuse the terminology and use a squared distance in our definition:

$$d\left(x,\tilde{w}\right)=\frac{1}{2N}\left\|x-\tilde{w}\right\|_2^2 \tag{S47}$$

which can be computed from the BP messages at convergence, as follows. First define the auxiliary quantities (representing the cavity variance and mean of each variable without the u_D field):

$$r_{ki\to D}=\left\langle x_{ki}^2\right\rangle_{h_{ki\to D}}-\left\langle x_{ki}\right\rangle_{h_{ki\to D}}^2=\left(\zeta_{ki}-\lambda\right)^{-1} \tag{S48}$$

$$\mu_{ki\to D}=\left\langle x_{ki}\right\rangle_{h_{ki\to D}}=\frac{\tilde{\mu}_{ki}-\lambda\tilde{w}_{ki}}{r_{ki\to D}} \tag{S49}$$

Then the expression of the average distance reads:

$$\begin{aligned}\left\langle d\left(x,\tilde{W}\right)\right\rangle&=\frac{1}{N}\sum_{ki}\frac{\int \mathrm{d}x_{ki}\,u_D\left(x_{ki}\right)h_{ki\to D}\left(x_{ki}\right)\frac{1}{2}\left(x_{ki}-\tilde{w}_{ki}\right)^2}{\int \mathrm{d}x_{ki}\,u_D\left(x_{ki}\right)h_{ki\to D}\left(x_{ki}\right)}\\&=\frac{1}{N}\sum_{ki}\frac{1}{2\lambda}\left(\frac{r_{ki\to D}}{r_{ki\to D}+\lambda^{-1}}+\frac{1}{\gamma}\left(\frac{\mu_{ki\to D}-\tilde{w}_{ki}}{r_{ki\to D}+\lambda^{-1}}\right)^2\right)\end{aligned} \tag{S50}$$

The free entropy of the system can be written as before, eq. (S42), but now in the computation of the entropy we also need to account for the energy of the extra field:

$$S=\phi+\frac{\chi+\lambda}{2}-\frac{\lambda}{N}\sum_{ki}\tilde{w}_{ki}\frac{\tilde{\mu}_{ki}}{\lambda_{ki}}-\log\left(\sqrt{2\pi e}\right) \tag{S51}$$

By varying λ and using eqs. (S50) and (S51), we can obtain a plot of the local entropy as a function of the distance around any given configuration $\tilde{W}$, as long as the BP equations converge. As a general rule of thumb, the equations don't converge when λ is too low in the 1-RSB phase, and when λ is too large and $\tilde{W}$ is not a solution. The equations do converge however even in the 1-RSB phase for large enough λ if $\tilde{W}$ is a solution, which can be understood as the external field breaking the symmetry. When $\tilde{W}$ is not a solution, going to the limit $\lambda\to\infty$ eventually restricts the BP equations to a region of the configuration space without solutions, leading to non-normalizable messages (this could of course be amended e.g. by just working at non-zero temperature). Besides these situations, other numerical problems may arise under certain circumstances when N is not large enough, due to the approximations in the messages.

As a consistency check, it can be verified numerically that formula (S51) yields the expected result for the $\alpha=0$ case, $S=\log\left(\sqrt{D\left(2-D\right)}\right)$.

C. Focusing-BP

In order to implement the Focusing-BP protocol we need to add a new type of node to each variable. These nodes do not directly represent energy terms in the usual sense, and therefore they are not factor nodes; rather, their role is that of effectively representing an interaction of y identical replicas of the original system with an extra auxiliary configuration $x^\star$. The derivation (in a discrete setting) can be found in ref. [5].

We denote as $u_{\star\to ki}\left(x_{ki}\right)$ the new extra field to be multiplied in eq. (S19). This field, like all others over the x_{ki} variables, is parametrized by two quantities $U_{\star\to ki}$ and $V_{\star\to ki}$ which get added to $\tilde{\mu}_{ki}$ and ζ_{ki} (eqs. (S31) and (S32)).The update equation is rather involved, but it can be simplified by breaking it down in steps and adopting a new notation for messages composition, leading to this expression:

$$u_{\star\to ki}=\left(\left(\left(h_{ki\to\star}*\lambda\right)\uparrow\left(y-1\right)\right)\otimes u_{\mathrm{n}^\star}\right)*\lambda \tag{S52}$$

where the notation is as follows:

- $u=h*\lambda$ represents (intuitively speaking) the effect of passing a message h through a Gaussian interaction with strength λ. In formulas:

$$u\left(\tilde{x}\right)=\int \mathrm{d}x\,e^{-\frac{\lambda}{2}\left(\tilde{x}-x\right)^2}h\left(x\right)$$

- $h = u_1 \otimes u_2$ represents the "composition" of two messages, in formulas:

$$h(x) \propto u_1(x)\, u_2(x)$$

which, using the internal representation in terms of rescaled mean $\tilde{\mu}$ and inverse variance ζ for the h, and the corresponding quantities for the $u_{1/2}$ fields, simply translates to:

$$\tilde{\mu} = U_1 + U_2$$
$$\zeta = V_1 + V_2$$

- $\hat{u} = u \uparrow y$ with integer y represents the composition of u with itself (i.e. $\hat{u} = u \otimes u \otimes \cdots \otimes u$) y times. This can be trivially extended to non-integer y and yields:

$$\hat{U} = yU$$
$$\hat{V} = yV$$

- $u_{\mathrm{n}^\star}$ is a normalization field, required for normalization of the $x^\star$ variables analogously to the u_{n} field for the x variables; analogously to that field, it is parametrized with a single parameter $\chi^\star$ and can be represented as $U_{\mathrm{n}^\star} = 0$, $V_{\mathrm{n}^\star} = \chi^\star$ (cf. equations (S19), (S20) and (S31)).

- $h_{ki\to\star}$ is the cavity field from a variable x_{ki} to the corresponding variable $x^\star_{ki}$, which can be defined from the relation $m_{ki} = h_{ki\to\star} \otimes u_{\star\to ki}$.

Therefore formula (S52) can be intuitively understood as follows: the cavity messages coming from the replicated variable nodes ($h_{ki\to\star}$) are passed through their interaction (with strength λ) with the corresponding variable $x^\star_{ki}$; the resulting messages are composed together (there are $y-1$ identical messages) and with the normalization field $u_{\mathrm{n}^\star}$; the resulting cavity message $h_{\star\to ki}$ is again passed through the interaction λ to yield the field $u_{\star\to ki}$.

The parameter $\chi^\star$ must be set to a value that normalizes the $x^\star$ variables; the norm of the $x^\star$ variables can be obtained from their marginals with the same formula used for the x, eq. (S39); the marginals can be obtained with this formula:

$$m_\star = ((h_{ki\to\star} * \lambda) \uparrow y) \otimes u_{\mathrm{n}^\star}$$

V. EXPERIMENTS WITH RANDOMIZED FASHION-MNIST

Our protocol to produce a randomized versions of the Fashion-MNIST dataset was as follows. Starting with the binarized, two-class dataset described in the main text, new input patterns were derived from the original ones as $\xi_i^{\mu\prime} = \xi_i^{\pi_i(\mu)}$ where $\pi_i(\cdot)$ is a random permutation, different for each i (i.e. we shuffled each pixel across samples). In this way, each pixel i has the same bias as in the original dataset, $\mathbb{E}_\mu[\xi_i'] = \mathbb{E}_\mu[\xi_i]$, but the (connected) correlations are destroyed, $\mathbb{E}_\mu[\xi_i'\xi_j'] \approx \mathbb{E}_\mu[\xi_i']\,\mathbb{E}_\mu[\xi_j']$, and furthermore the patterns $\xi^{\mu\prime}$ no longer carry information about the target label σ^μ, so that no generalization is possible. As a result, the randomized patterns carry more information per pixel that needs to be stored by the device, which in turn can be expected to make the learning problem harder.

We produced 5 such shuffled datasets and performed 10 tests on each with the same algorithms as for the Fashion-MNIST tests, using the same parameters when possible. The exception was eLAL, which we had to tweak slightly to avoid divergencies: we used $\eta = 5 \cdot 10^{-4}$, $\lambda_0 = 20$, $\lambda_1 = 1 \cdot 10^{-2}$.

We directly compared the results with those obtained on the original dataset. Fig. S3 shows that the maximum eigenvalues (cf. fig. 6 of the main text, bottom panel) hardly change between the two tests, with only a slight degradation for the eLAL algorithm.

Fig. S4 shows the volumes around the solutions, computed with the BP algorithm; this is the analogous of fig. 6 of the main text, top panel. We also kept the same scale for easiness of comparison. Again, we observe a slight degradation of the eLAL algorithm compared to ceSGD, albeit only at short distances. It is still the case that LAL is by far the worst algorithm, and that ceSGD slow is better than ceSGD fast. The volumes are overall smaller than for the original dataset.

We also performed an additional experiment on both datasets, measuring the robustness towards the presence of noise in the input. For each input image, the noise was added by replacing a randomly selected fraction η of pixels with random binary values. Each new value of a pixel i was extracted with the same bias observed in the original dataset for that pixel, $\mathbb{E}_\mu[\xi_i]$: in this way, the new corrupted images were still rather close to the original input

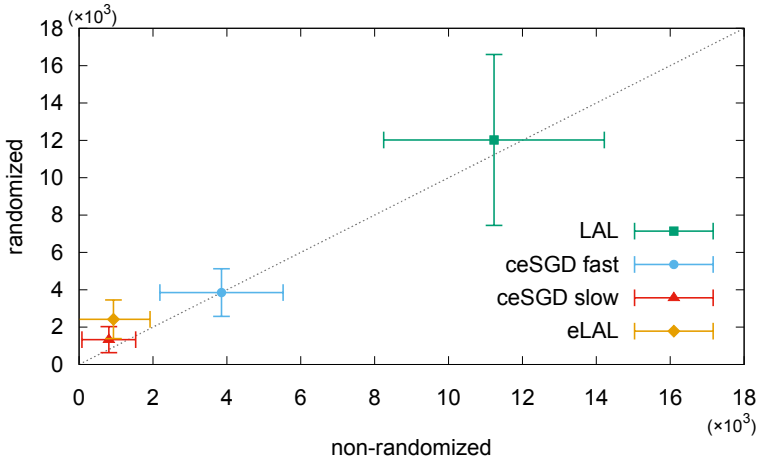

Figure S3. Comparison of the distributions of the maximum eigenvalues (means and stadard deviations) between the original Fashion-MNIST dataset and its randomized version, for various algorithms. The "non-randomized" values use the same data as the clouds of points of fig. 6 of the main text, bottom panel).

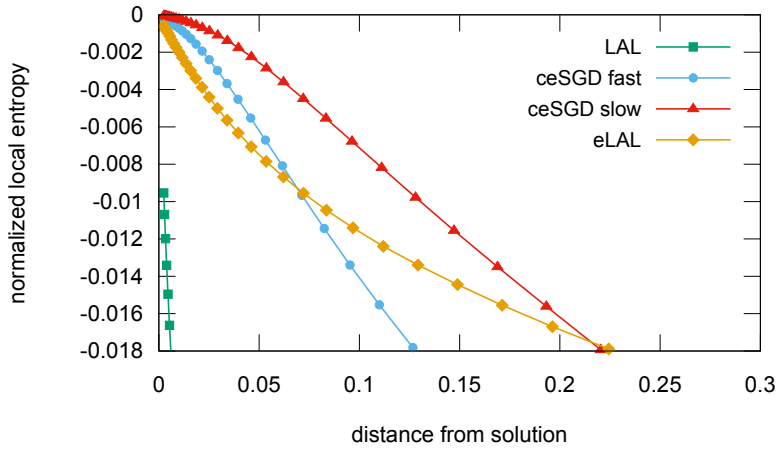

Figure S4. Normalized local entropies as a function of the distance for solutions found by various algorithms, using the randomized-Fashion-MNIST dataset. This is the analogous of fig. 6 of the main text, top panel, on the randomized dataset.

distribution (although the correlation with the desired output label, and the internal correlations between pixels, were degraded) and the networks would be able to operate in the same regime in which they were trained. This model of noise is intended to provide a rough proxy for the generalization capabilities of the networks without the need for a validation set, since it measures the amount of overfitting within a manifold that should approximate the distribution of the data to be classified. The results are shown in fig. S5, and they are fully consistent with the picture emerging from the study of the local volumes.

[1] 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. *Physical Review Letters*, 115(12):128101, September 2015. ISSN 0031-9007, 1079-7114. doi:10.1103/PhysRevLett.115.128101.

[2] Rémi Monasson. Structural glass transition and the entropy of the metastable states. *Physical review letters*, 75(15):2847, 1995. doi:10.1103/PhysRevLett.75.2847.

[3] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001. doi:10.1017/CBO9781139164542.

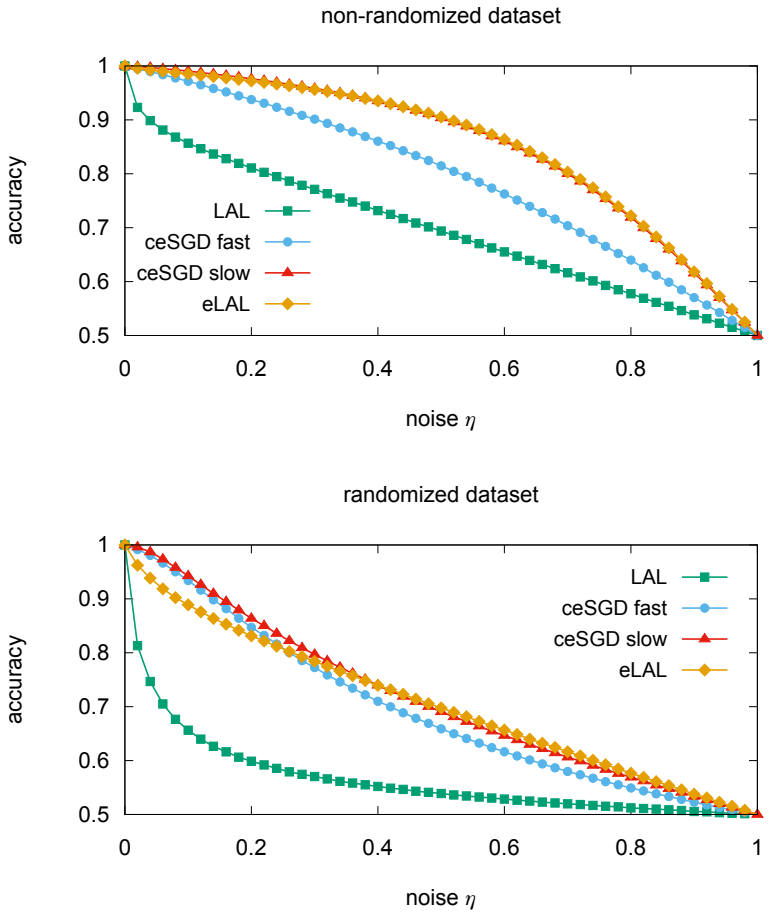

Figure S5. Training accuracy in the presence of (biased) noise, for the original Fashion-MNIST dataset (top panel) and for its randomized version (bottom panel). In the top panel, eLAL and ceSGD slow are hardly distinguishable.

[4] 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.

[5] 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. ISSN 0027-8424, 1091-6490. doi:10.1073/pnas.1608103113.

[6] 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.

[7] 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.

[8] Florent Krzakala, Federico Ricci-Tersenghi, Lenka Zdeborova, Riccardo Zecchina, Eric W Tramel, and Leticia F Cugliandolo. *Statistical Physics, Optimization, Inference, and Message-Passing Algorithms: Lecture Notes of the Les Houches School of Physics-Special Issue, October 2013*. Number 2013. Oxford University Press, 2016. doi:10.1093/acprof:oso/9780198743736.001.0001.

[9] Alfredo Braunstein and Riccardo Zecchina. Learning by message passing in networks of discrete synapses. *Physical Review Letters*, 96:030201, Jan 2006. doi:10.1103/PhysRevLett.96.030201.