Wide flat minima and optimal generalization in classifying high-dimensional Gaussian mixtures

Wide flat minima and optimal generalization in classifying high-dimensional Gaussian mixtures

Carlo Baldassi,[1] Enrico M. Malatesta,[1] Matteo Negri,[1,2] and Riccardo Zecchina[1]

[1] *Artificial Intelligence Lab, Bocconi University, 20136 Milano, Italy*
[2] *Dept. Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy*

We analyze the connection between minimizers with good generalizing properties and high local entropy regions of a threshold-linear classifier in Gaussian mixtures with the mean squared error loss function. We show that there exist configurations that achieve the Bayes-optimal generalization error, even in the case of unbalanced clusters. We explore analytically the error-counting loss landscape in the vicinity of a Bayes-optimal solution, and show that the closer we get to such configurations, the higher the local entropy, implying that the Bayes-optimal solution lays inside a wide flat region. We also consider the algorithmically relevant case of targeting wide flat minima of the (differentiable) mean squared error loss. Our analytical and numerical results show not only that in the balanced case the dependence on the norm of the weights is mild, but also, in the unbalanced case, that the performances can be improved.

CONTENTS

I. INTRODUCTION

Some basic aspects of contemporary machine learning (ML) do not find a satisfactory explanation in the classical statistical learning framework. For instance, the over-parametrized regime in which deep neural networks (DNN) are utilized does not easily fit into the uniform convergence scenario in which one expects that the complexity of a machine learning device (function) should be of the order of the number of examples to provide good generalization properties [1]. In order to make progress, researchers are trying to connect the generalization capabilities of devices such as DNN with the geometrical properties of the error loss functions that is minimized for learning. An interesting conjecture which has emerged in various contexts argues that the flatness of the minima can lead to good generalization in the over-parametrized regime [2–6]. The fact that such minima have been recently shown to exist already in shallow networks [3, 7–9], puts this conjecture on solid grounds. Results based on statistical physics techniques, have shown that there exist relatively wide regions with a very high density of optimal minimizers. These regions coexist with a much larger number of critical points (narrow local minima and saddles) and are called either high local entropy regions or wide flat minima.

So far, the theoretical results were limited to the so-called teacher-student scenario in the context of classification: a training set with i.i.d. randomly-generated inputs and labels provided by a (shallow) teacher network is presented to a student network with the same architecture as the teacher. In the over-parametrized regime, in which the training set does not contain sufficient information, several local minima exist, and the ones with high local entropy were shown to have generalization capabilities close to the Bayesian one.

The over-parametrized teacher-student scenario considered in the above-mentioned studies is highly non-convex, and using random i.i.d inputs is not necessarily realistic. Although it was shown in [7] that the phenomenology is similar with real datasets, the problem of obtaining theoretical insight into other classification tasks with different distributions remains open. Some preliminary results in this direction are given in [10–14].

Here, we discuss the connection between local entropy, flatness and generalization in a very basic model of high-dimensional statistical machine learning [15–18]: Gaussian mixtures. The generative model is defined as follows. For a given problem size N, an N-dimensional vector $\mathbf{v}^\star$ is randomly generated from a standard multivariate normal $\mathcal{N}(\mathbf{0}, \mathrm{I}_N)$. Then we generate a label $\sigma = 1$ or $\sigma = -1$ with probability ρ and $1-\rho$, respectively, and we generate a pattern $\boldsymbol{\xi}$ according to the value of the label σ as $\mathcal{N}\left(\sigma \mathbf{v}^\star/\sqrt{N}, \Delta\, \mathrm{I}_N\right)$. In this way we construct a training set with $P \equiv \alpha N$ such patterns; the coordinate $i \in \{1, ..., N\}$ of pattern μ is therefore given by

$$\xi_i^\mu = \frac{v_i^\star}{\sqrt{N}}\sigma^\mu + \sqrt{\Delta}\, z_i^\mu \tag{1}$$

where z_i^μ are i.i.d Gaussian random variables with zero mean and unit variance. This results in two potentially overlapping clusters, with the label indicating the cluster a pattern belongs to and where Δ controls their width. We will refer to the problem with $\rho = 0.5$ as the *balanced* case; we call all other cases *unbalanced*.

As usual in statistical physics, we will consider the high-dimensional limit, where both $N \to \infty$ and $P \to \infty$ with the ratio $\alpha = {}^P\!/_N$ fixed.

In this paper we analyze the performance of a threshold-linear classifier (a single-unit neural network). This machine is parametrized by a vector of weights $\boldsymbol{w}$ of length N and a bias b, but when studying the balanced case we always simply set $b = 0$. The machine predicts the label of a pattern $\boldsymbol{\xi}$ as:

$$\hat{\sigma}\left(\boldsymbol{\xi}; \boldsymbol{w}, b\right) = \mathrm{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i\, \xi_i + b\right) \tag{2}$$

We can observe that this classifier is invariant to an overall rescaling of the parameters $\boldsymbol{w}' = \kappa \boldsymbol{w}$ and $b' = \kappa b$ for any $\kappa > 0$. It is natural to identify the irrelevant degree of freedom in the parametrization with the norm of $\boldsymbol{w}$.

The most elementary metric by which to measure the performance of this classifier on the training set is the number of errors, which can be expressed as

$$\mathscr{L}_{\mathrm{err}}(\boldsymbol{w}, b) = \sum_{\mu=1}^{P} \Theta\left[-\sigma^\mu\, \hat{\sigma}(\boldsymbol{\xi}^\mu; \boldsymbol{w}, b)\right] \tag{3}$$

where $\Theta(\cdot)$ is the Heaviside step function, that is $\Theta(x) = 1$ if $x \geq 0$ and 0 otherwise. This error-counting loss obviously inherits the scale invariance, but it has the drawback that it cannot be used with the gradient-based methods usually employed in training large neural networks (which is the situation about which we hope to gain the most insight from

this simple model). It is therefore of interest to consider a generalized overall loss function form:

$$\mathscr{L}(\boldsymbol{w}, b) = \sum_{\mu=1}^{P} \ell \left[\sigma^{\mu} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \, \xi_i^{\mu} + b \right) \right] \tag{4}$$

where $\ell(\cdot)$ is a generic single-pattern loss function. The error-counting case corresponds to $\ell(x) = \Theta(-x)$. In what follows we analyze the mean squared error (MSE) loss $\ell(x) = \frac{1}{2}(x-1)^2$, which is a well-studied differentiable loss. As for other choices of differentiable losses (e.g. cross-entropy, hinge, etc.) the scale-invariance property is lost, and the role of norm regularizations may become important.

It's important to observe that this model possesses some rather peculiar features if compared to typical classification tasks performed with neural networks, namely the training loss is convex. Indeed, Bayes-optimal performance can be achieved with a single configuration of the model parameters (instead of requiring a distribution) which can be derived analytically. Additionally, due to the overlap between the Gaussians which are used to generate the data, no classifier can achieve zero test error (in the teacher-student context this would be somewhat similar to the case of having a "noisy", unreliable teacher). Therefore, care must be used when considering how the results may generalize to non-convex scenarios.

Recently, this model has been studied in [19] by using Gordon's inequality. The authors showed that the MSE loss is severely prone to overfitting, especially when $\alpha \simeq 1$[1]. They also showed however that, in spite of the fact that the output of the model is norm-independent, the generalization performance is considerably improved by adding to the loss an L_2 regularization term on the weights, $\lambda \|\boldsymbol{w}\|^2$. For large N, the parameter $\lambda > 0$ is a Lagrange multiplier that implicitly fixes the norm of the weights. The optimal choice of λ depends in general on α and ρ. For the balanced case, $\rho = 0.5$, the optimum is obtained for all α in the limit $\lambda \to \infty$ (corresponding to vanishing values for the norm of the weights), and in that case the network reaches the Bayes-optimal generalization bound.

In light of these findings, it is interesting to further discuss the role of the norm for these class of models. As we remarked above, the output of the network (and thus the generalization error) has a scale invariance, i.e. it is independent of the norm (as long as the bias is also properly rescaled). As a consequence we need to understand which are the geometrical features of the solutions space of the classifier that are induced by the regularization of the surrogate loss function used for gradient learning.

It is worth noticing that this scenario also applies to most deep neural network models that use ReLU activations in the intermediate layers and an argmax operation to produce the output label, and are therefore invariant to uniform scaling of all their weights and biases. Since the norm cannot affect the generalization capabilities of the network, it seems unlikely that a norm-based regularization could be a valid general strategy.[2]

In this paper, we argue for a different, more general criterion to avoid overfitting and improve generalization, proposed in several recent works [7, 9, 20], namely that of maximizing the local entropy, which is a particular measure of flatness that can be analyzed theoretically and efficiently approximated algorithmically. We refer to gradient-based algorithms that operate by maximizing (an approximation to) the local entropy as "entropic algorithms". One example is given by the replicated stochastic gradient descent (rSGD) algorithm introduced in [8], where the local entropy is targeted by using several replicas moving in the loss landscape and at the same time feeling an attraction during their dynamics. Another algorithm is entropy-SGD (eSGD) [21], where the local entropy is estimated using stochastic gradient Langevin dynamics [22]. Those algorithms have been applied to state of art deep neural networks [23], proving that they can achieve improved generalization performances.

In particular, for the balanced case, we show analytically that the minimum norm condition, which results in Bayes-optimal performance, corresponds to solutions of maximum local entropy for the classifier (which is norm invariant). We also show that these solutions can be found by entropic algorithms acting on the MSE loss function, and that these algorithms are much less sensitive to the norm.

For the unbalanced case, the authors of [19] found that, when the bias is learned, reaching the Bayes-optimal generalization error with L_2 regularization alone is impossible, and that there exists an optimal finite value of λ that minimizes the generalization error. In this paper we show however that that there exists a choice of the bias and of λ that allows to reach the Bayes-optimal performance, and that such parameters also have a higher local entropy (measured again in a scale-invariant way). We show both analytically and numerically how to systematically improve the generalization performance in this setting. Learning the bias with entropic algorithms leads to improved performance compared to those which can be attained by the L_2 regularized loss function.

---

[1] This value corresponds to the transition point above which the MSE loss has a unique minimum, since minimizing the MSE entails solving a system of P equations in N unknowns; in the balanced case, it is also the transition point where the data is no longer linearly separable.

[2] There is a caveat to this statement: for particular choices of the loss, e.g. cross-entropy, it is possible to reparametrize the problem in an invariant way and interpret the norm in terms of a time-evolving parameter of the loss with a "focusing" role, see [7].

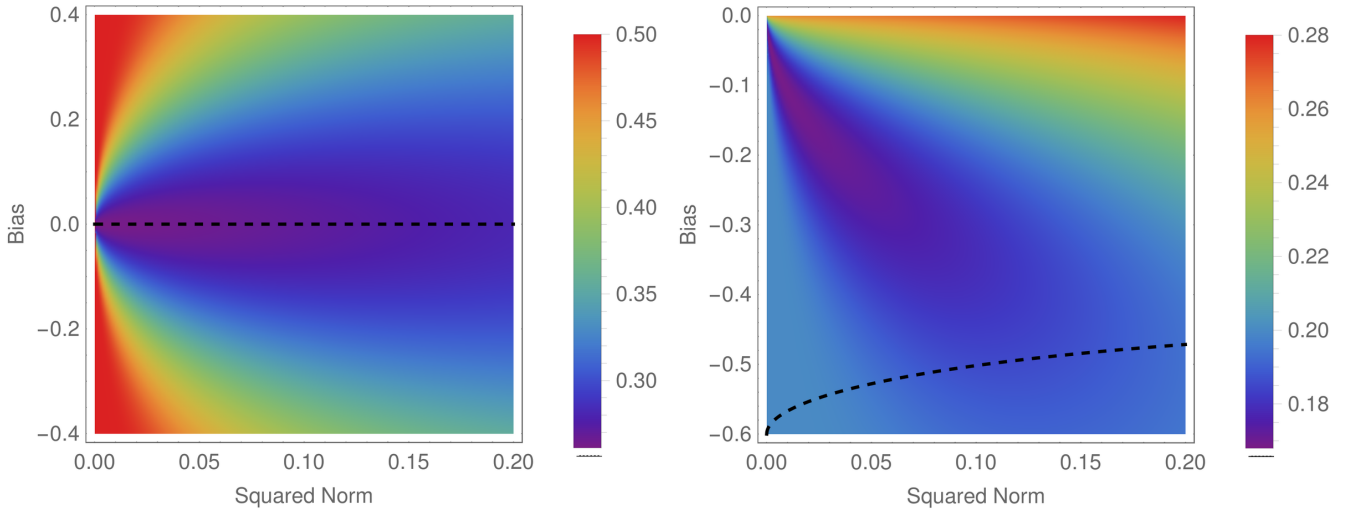

Figure 1. Generalization error found by optimizing the regularized MSE loss, as a function of the bias and squared norm. The black dashed line represents the value of the bias obtained by maximizing the Gardner volume. Both plots are for $\alpha = 0.7$ and $\Delta = 1$. Left: Balanced case ($\rho = 0.5$). The Bayes optimal generalization error in this case is $\epsilon_g \simeq 0.2605\ldots$ computed using equation (F6) of appendix F. Right: Unbalanced case, with $\rho = 0.2$. The Bayes optimal generalization error in this case is $\epsilon_g \simeq 0.1679\ldots$.

The rest of the paper is organized as follows. In section II we briefly review the typical scenario obtained by performing a standard replica-symmetric (RS) replica calculation over the Gibbs measure. We discuss in particular how the choice of the bias is decisive for generalization performances in the unbalanced case. In section III we explore the local entropy landscape around the Bayes optimal configuration for the MSE loss function, by performing a calculation á la Franz-Parisi [24]. In section IV we discuss how targeting the local entropy loss we can improve generalization. Finally section V contains some conclusions.

II. BAYES-OPTIMAL CONFIGURATIONS

The partition function of the Gaussian mixture model, with a regularization over the weights of the linear classifier can be written as

$$Z = \int \prod_i dw_i \, e^{-\beta \mathscr{L}(\boldsymbol{w},b) - \frac{\lambda}{2}\sum_i w_i^2} \tag{5}$$

where β is the inverse temperature. Notice that in our treatment the bias has been fixed as an external parameter. To study the case in which the bias is a learned parameter we would add another integral over b in the definition of Z. In the following we will denote the average over the training set with angle brackets: $\langle\cdot\rangle \equiv \prod_{\mu=1}^{P} \mathbb{E}_{\boldsymbol{v}^\star,\sigma^\mu} \mathbb{E}_{\boldsymbol{\xi}^\mu|\boldsymbol{v}^\star,\sigma^\mu} [\,\cdot\,]$.

The typical properties of the model are derived by computing the average log-volume $\langle \ln Z \rangle / N$, which is the (typical) free entropy of the model $-\beta f$, where f is the corresponding free energy. The free entropy can be computed in the large-N limit using the "replica trick":

$$\ln Z = \lim_{n\to 0} \partial_n Z^n \,. \tag{6}$$

The whole computation in the large N and β limit using an RS ansatz is reported in appendix A and B. Here we discuss the results. In figure 1 we show the generalization error found by optimizing the regularized MSE loss function, in the balanced case and one unbalanced case, as a function of the bias and the squared norm (which is implicitly but monotonically controlled by λ). We also show with a black dashed line the value that the bias takes when it is learned, for any given value of the squared norm. In both the balanced and unbalanced cases there exist choices of the bias and the squared norm that achieve the Bayes optimal performance. However, an important difference can be noted: if we learn the bias in the balanced case we always find $b = 0$ for every value of the squared norm; sending $\lambda \to \infty$ (and therefore the squared norm to zero) one recovers the Bayes optimal performance. This is not true in the unbalanced case: fixing λ and learning the bias never gives the optimal performance.

III. THE LOCAL ENTROPY IS LARGER IN THE VICINITY OF THE BAYES-OPTIMAL CONFIGURATION

In order to quantify the local geometrical landscape around a typical configuration $\tilde{\boldsymbol{w}}$ of the Gibbs measure with loss function $\mathscr{L}_r = \sum_\mu \ell_r$, regularization parameter λ_r and inverse temperature β_r, we have studied the so-called Franz-Parisi free entropy [24, 25]. It is defined as

$$-\beta f_{\rm FP}(S) \equiv \frac{1}{N}\left\langle \frac{\int \prod_i d\tilde{w}_i \, e^{-\beta_r \mathscr{L}_r(\tilde{\boldsymbol{w}},\tilde{b}) - \frac{\lambda_r}{2}\sum_i \tilde{w}_i^2} \ln \mathcal{V}\left(\tilde{\boldsymbol{w}}, S\right)}{\int \prod_i d\tilde{w}_i \, e^{-\beta_r \mathscr{L}_r(\tilde{\boldsymbol{w}},\tilde{b}) - \frac{\lambda_r}{2}\sum_i \tilde{w}_i^2}} \right\rangle \tag{7}$$

where the quantity

$$\mathcal{V}\left(\tilde{\boldsymbol{w}}, S\right) \equiv \int d\mu_P(\boldsymbol{w}) \, e^{-\beta \mathscr{L}(\boldsymbol{w},\tilde{b})} \delta\left(\sum_i w_i \tilde{w}_i - NS\right) \tag{8}$$

is the volume of configurations $\boldsymbol{w}$ at inverse temperature β that have overlap S with the reference configuration $\tilde{\boldsymbol{w}}$. The measure $d\mu_P(\boldsymbol{w})$ is a flat measure over a hyper-sphere of squared radius P, i.e. the weights $\boldsymbol{w}$ have square norm $\frac{1}{N}\sum_i w_i^2 = P$. The overlap P is chosen to match the squared norm of the reference $\tilde{\boldsymbol{w}}$, that is $P = Q$. Note that Q is fixed via a soft constraint by the regularization parameter λ_r; in addition we have chosen, for simplicity, the bias of the constrained configuration $\boldsymbol{w}$ to match the one of the reference.

Notice that in eqs. (7) and (8) we use different losses $\mathscr{L}_r$ and $\mathscr{L}$ (and different parameters too): the landscape of which we explore the geometrical features can differ from the landscape from which we get the reference configuration.

The computation of eq. (7) is long and involved; here we just sketch the main steps, referring to the appendix C for the details. The average over the disorder in eq. (7) can be done by using two replica tricks, one for the denominator and another one for the logarithm in the numerator:

$$\frac{1}{Z} = \lim_{r\to 0} Z^{r-1} \tag{9a}$$

$$\ln Z = \lim_{n\to 0} \partial_n Z^n \tag{9b}$$

Once the average is performed, one has to introduce several order parameters in order to decouple the expressions over the size of the training set αN and of the dimension N. Using indexes a or b for replicas in $\{1,\dots,r\}$ and c, $d \in \{1,\dots,n\}$ the order parameters are $p^{cd} = \frac{1}{N}\sum_i w_i^c w_i^d$, $t^{ac} = \frac{1}{N}\sum_i \tilde{w}_i^a w_i^c$, $O^c = \frac{1}{N}\sum_i v_i^\star w_i^c$, $P^c = \frac{1}{N}\sum_i (w_i^c)^2$ and the corresponding conjugated ones. Note that P^c is just the squared norm P because of the spherical constraint inside the measure $d\mu_P(\boldsymbol{w})$. Among the conjugated order parameters, we also need to introduce an additional parameter, $\hat{S}^c$, which imposes the hard constraint on the overlap between the reference configuration $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$.

Using an RS ansatz over the order parameters (see appendix C 1) and performing the large β_r limit we obtain

$$-\beta f_{\rm FP}\left(S\right) = \mathfrak{G}_S + \alpha \mathfrak{G}_E \,, \tag{10}$$

where the definition of the entropic and energetic terms are reported in appendix C 2. When $\alpha = 0$ the Franz-Parisi free entropy can be evaluated analytically

$$-\beta f_{\rm FP}\left(S, \alpha = 0\right) = \frac{1}{2}\left[1 + \ln\left(2\pi\right) + \ln\left(\frac{1}{\lambda_r} - \lambda_r S^2\right)\right] , \tag{11}$$

and it gives the logarithm of the total volume of configurations at overlap S with the reference. For a given loss $\ell\left(\cdot\right)$ the local entropy of a given configuration $\tilde{\boldsymbol{w}}$ can be computed by evaluating the local energy $\epsilon_\ell = \frac{\partial(\beta f_{\rm FP})}{\partial \beta}$ and then using

$$\mathcal{S} = \beta\left(\epsilon_\ell - f_{\rm FP}\right) . \tag{12}$$

The normalized local entropy is just the local entropy (12) minus the total log-volume at $\alpha = 0$ given in equation (11). It is the logarithm of a fraction of a volume, and thus is upper bounded by zero; additionally, defining the distance as

$$d \equiv \frac{1}{2}\frac{\sum_{i=1}^N \left(\tilde{w}_i - w_i\right)^2}{\sum_{i=1}^N \tilde{w}_i^2} = 1 - \frac{S}{P} \tag{13}$$

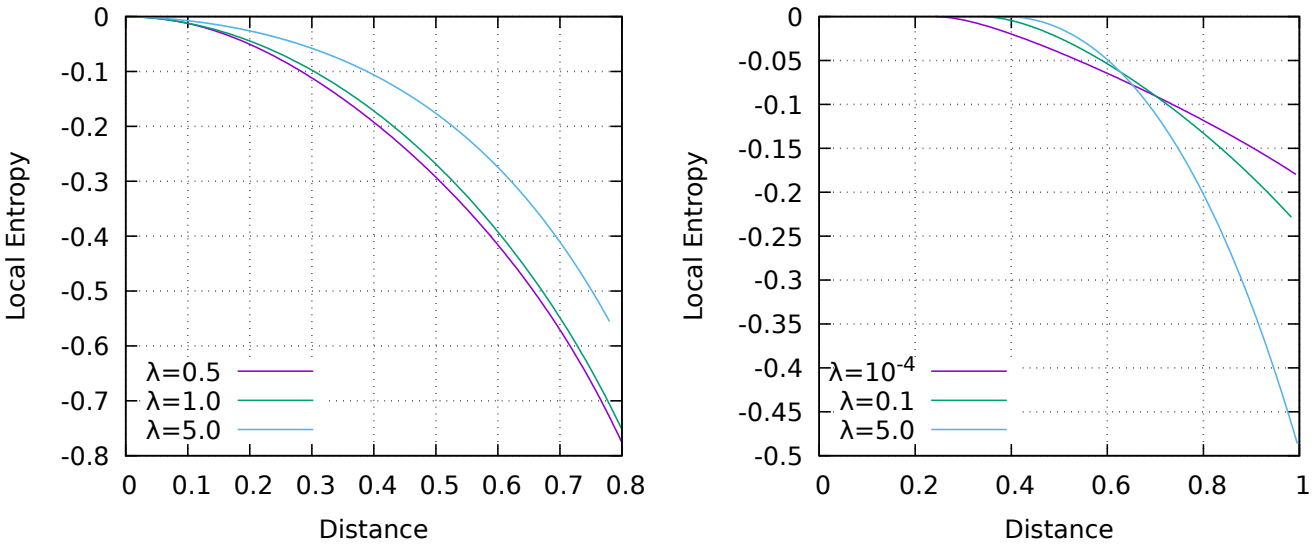

Figure 2. Balanced case ($\rho = 0.5$). Normalized local entropy as a function of the distance d computed from reference configurations found by optimizing the regularized MSE loss, with varying regularization strength λ. Larger values of λ correspond to minimizers with better generalization properties. In both figures $\alpha = 0.7$, $\Delta = 1$, $b = 0$. The cutoff $\bar{\epsilon}$ is chosen to be equal to the training error of the reference (left), or is given by the training error of the "oracle" $\boldsymbol{w} = v^\star$ (right panel) as in equation (14).

the normalized local entropy is always zero for $d = 0$. For $\tilde{\boldsymbol{w}}$ located in sharp minima we expect that the normalized local entropy will have a sharp drop near $d \simeq 0$, whereas for flat minima it will be close to zero within some range of small distances.

We have explored the normalized local entropy landscape of the configurations found by optimizing the regularized MSE loss (i.e. $\ell_r(x) = \frac{1}{2}(x-1)^2$), where the *training error* is used in the local entropy definition (i.e. $\ell(x) = \Theta(-x)$). We stress that by using the error instead of the MSE we explore the properties of the model in the regime in which it operates during classification.

On the other hand, the parameter β has been chosen in such a way that the training error of $\boldsymbol{w}$ given in (C11) is equal to a certain cutoff $\bar{\epsilon}$.

We have analyzed two different choices for the energy $\bar{\epsilon}$:

- in the first case $\bar{\epsilon}$ is chosen to be equal to the training error of the reference (see eq. (B16) in the appendix). This case is depicted in the left panel of figure 2 for the balanced case and in the right panel of figure 3 for the unbalanced case. See the corresponding captions for details.

- in the second case, only used for the balanced case, $\bar{\epsilon}$ is chosen as the training error of an "oracle classifier" with $\boldsymbol{w} = v^\star$, which is given by:

$$\epsilon_t^\star = \alpha \int \prod_i dv_i^\star d\xi_i P(\xi_i \mid v_i^\star) P_v(v_i^\star)\, \theta\left(-\frac{1}{\sqrt{N}} \sum_i v_i^\star \xi_i\right) = \alpha H\left(-\frac{1}{\sqrt{\Delta}}\right) \tag{14}$$

  This corresponds to the smallest possible test error that any linear classifier machine could achieve, which is non-zero because of the overlap between the two clusters. This case is depicted in the right panel of figure 2.

In both cases we clearly see that reference configurations with better generalization properties have higher local entropy curves. We remind that, in the balanced case, the configurations with better generalization properties correspond to larger values of the regularization parameter λ, whereas in the unbalanced case they correspond to particular fine-tuned values of the bias b and λ as already evidenced in the right panel of figure 1.

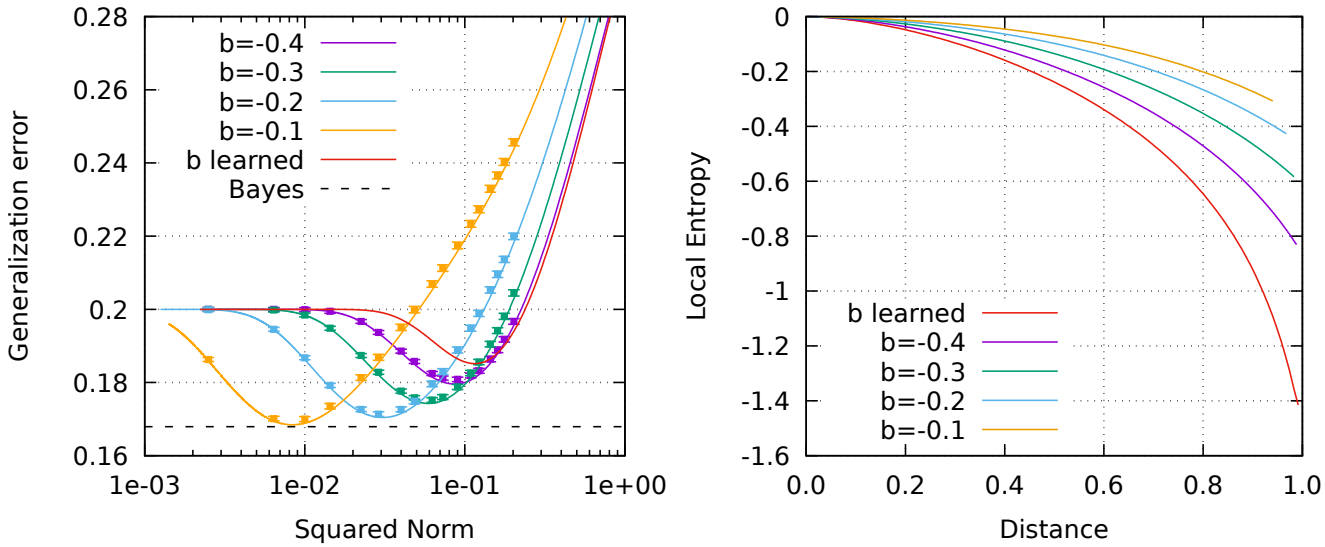

Figure 3. Unbalanced case ($\rho = 0.2$). In both plots $\alpha = 0.7$ and $\Delta = 1$. Left: Generalization error for a typical minimizer of the MSE loss as a function of the squared norm, for various choices of the bias b = -0.4, -0.3, -0.2, -0.1. Full lines are analytical results, points and error bars are numerical results obtained with $N = 1000$. On the red curve, instead, the bias is learned (it's the value that maximizes the free entropy) and thus it's different for every value of the squared norm. The dashed curve is the Bayesian generalization error (see eq. (F6) in the appendix F). Right: Normalized local entropy as a function of the squared-distance d computed from reference configurations found by optimizing the regularized MSE loss, for various choices of the bias b_r = -0.4, -0.3, -0.2, -0.1 (and $b = b_r$). The corresponding value of the squared norm has been chosen by using the one that minimizes the generalization error for that fixed value of b (see left panel). In the red curve, instead, the bias has been fixed by a saddle point equation (i.e. it is learned). The cutoff $\bar{\epsilon}$ is chosen to be equal to the training error of the reference.

IV. ALGORITHMS THAT TARGET FLATTER REGIONS OF THE MSE LANDSCAPE ALSO GENERALIZE BETTER

The results of the previous section confirm that the local entropy landscape constructed using the training error is a good predictor of generalization performance. However, when dealing with much more complex architectures, using the training error as the loss function in eq. (12) is not (yet) algorithmically feasible. In particular, the entropic algorithms rSGD and eSGD must still operate on a differentiable loss. This leaves the question whether targeting high-local-entropy regions in a differentiable loss landscape can still lead to good generalization results open. We have investigated this question analytically on the Gaussian mixture model with a linear classifier and the MSE loss, using the same technique explained in [7, 9], see appendix D for details. This amounts to studying the generalization error of the barycenter of a replicated system of y classifiers, each with its own parameters $\boldsymbol{w}^a$ with $a = 1, \dots, y$, each optimizing the MSE under constraints on their norms n and on their mutual angles θ, that is: $\forall a, a' : \ \|\boldsymbol{w}^a\| = n, \ \boldsymbol{w}^a \cdot \boldsymbol{w}^{a'} = n^2 \cos(\theta)$. The barycenter is defined as $\bar{\boldsymbol{w}} = \frac{1}{y} \sum_a \boldsymbol{w}^a$. In this analysis we used the angle θ rather than the distance in order to compare situations with different norms.[3] Notice also that we have not imposed analytically an analogous constraint over the biases of the replicas. In other words, the bias of every replica is the same and we call it b. In order to set the bias for the barycenter we recall the fact that the error-counting loss is scale invariant, so that the value of the bias is significant only when compared to the norm of the weights. Additionally we note that in general $\|\boldsymbol{w}^a\| \geq \|\bar{\boldsymbol{w}}\|$, so if we were to naively use b as bias of the barycenter we would change its relative magnitude respect to $\|\bar{\boldsymbol{w}}\|$. For this reason we set $\bar{b} = b \, \|\bar{\boldsymbol{w}}\| \, / \, \|\boldsymbol{w}\|$.

Our goal is to check if we can improve the generalization performances. Due to the peculiarities of this model, in the balanced case we can simply check whether the barycenter is aligned with the solution of the norm-regularized model with large λ, which we know to be the optimal classifier.

Some representative results are shown in fig. 4. In the left panel we analyze the balanced case. Our results indicate that, with sufficiently many replicas (even just $y = 3$) and with sufficiently large angles the generalization performance

[3] This is equivalent to using the cosine similarity, often employed in machine learning contexts. We should note however that in a multi-layer classifier the structure of the scale invariance is more complicated and the cosine similarity by itself would not be sufficient to account for it.

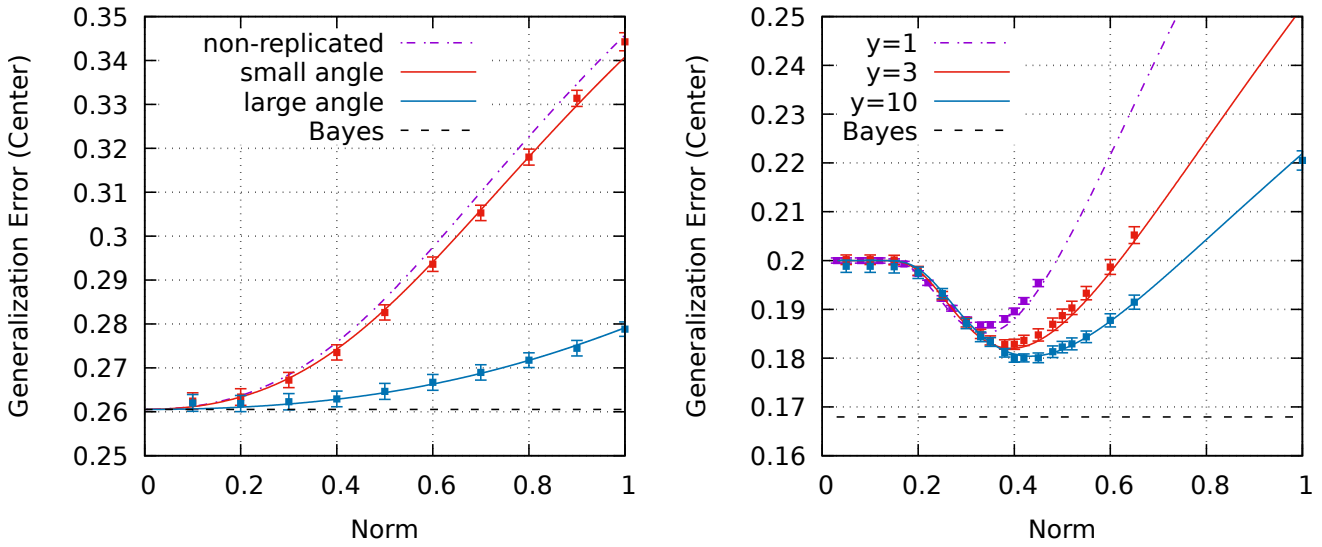

Figure 4. Generalization error of the center $\bar{\boldsymbol{w}}$ of a system of y replicas, each optimizing the MSE loss and with a constraint on the angle between the replicas, as a function of the norm n of the replicas. In both figures $\alpha = 0.7$; the Bayes optimal error is plotted with a dashed black line. Left: Balanced case with $y = 10$ replicas. The red curve (small-angle) corresponds to $\cos(\theta) = 0.9$; the large-angle case to $\cos(\theta) = 0.1$. Solid curves are theoretical results, points are numerical results obtained with $N = 1000$, averaged over 30 samples. In the limit $\theta = 0$ the results reproduce those of a single device; increasing θ the dependence on the norm reduces (the curve flattens onto the Bayes-optimal dashed line in the limit $\theta = \pi/2$ and $y \to \infty$). Right: Unbalanced case ($\rho = 0.2$). Solid curves correspond to analytical results, points are numerical results obtained with $N = 1000$ and respectively 100, 30, 20 samples for $y = 1, 3, 10$. The angle between the y replicas has been fixed to $\theta = \pi/2$.

is nearly optimal and the dependence on the norm is mild, and much less pronounced than at small angles (the limit of zero angles reproduces the results of the norm-regularized analysis without replicas). The fact that for this model the best results are obtained with widely separated replicas is due to the simple nature of the problem, and we do not expect this phenomenon to just carry over as-is to the case of deep neural networks, where the landscape is non-convex and the structure of the symmetries is generally much more complex (both in terms of the scale invariance and of discrete permutation symmetries).

In the right panel of figure 4 we show also what happens in the unbalanced case. We have compared the performance of the typical minimizer of the norm-regularized MSE loss with the one of the replicated system. We show that by increasing the number of replicas keeping fixed the angle between them, the generalization performance is improved. The large y limit can be handled analytically, and it is indistinguishable from the results obtained with $y = 10$ replicas.

The analytical curves describe very well the numerical results that are obtained with rSGD (see for all the details appendix G and [8] where it was firstly introduced). The algorithm consists in training y replicas of a perceptron with and additional term $\mathscr{L}_{\rm d}$ in the loss function of each model proportional to the sum of distances from the other replicas; in order to force the replicas to stay at a given distance d_0, we modify this term by including d_0 as offset:

$$\mathscr{L}_{\rm d}^{(a)} = \sum_{b \neq a}^{y} (d_{ab} - d_0)^2 \,, \tag{15}$$

where the index a refers to the replica of which we are computing the loss.

V. CONCLUSIONS

We have presented an analytical study concerning the connection between local entropy and optimal generalization in the case of Gaussian mixtures. Configurations of weights that reach Bayes-optimal performance were shown to be located inside regions of high local entropy, i.e. in wide flat minima of the error-counting loss function. We have also shown that targeting the wide flat minima of the differentiable loss function used for gradient learning (e.g. MSE) is a viable algorithmic strategy. Work is in progress to extend these type of results to deeper architectures.

ACKNOWLEDGMENTS

We wish to thank Francesca Mignacco and Luca Saglietti for useful discussions.

Appendix A: The replicated partition function

In this first appendix we briefly review how the the geometry of the space of typical Gibbs configurations of the model can be studied using statistical physics techniques [26, 27], such as the replica method.
In the Gaussian mixture problem, the joint probability distribution of patterns, labels and the centroid is

$$P(\boldsymbol{\xi},\sigma,\boldsymbol{v}^\star)=P_\sigma(\sigma)\prod_{i=1}^{N}\left[P(\xi_i|\sigma,v_i^\star)P_v(v_i^\star)\right] \tag{A1}$$

where $P_v(v_i^\star)=\mathcal{N}(0,1)$ and the conditional probability $P(\xi_i|\sigma,v_i^*)=\mathcal{N}\left(\frac{\sigma v_i^\star}{\sqrt{N}},\Delta\right)$, are Gaussian random variables and $P_\sigma(\sigma)=\rho\delta(\sigma-1)+(1-\rho)\delta(\sigma+1)$. As stated in the main text, we will denote the average over the disorder distribution given in equation (A1) with $\langle\cdot\rangle\equiv\prod_{\mu=1}^{P}\mathbb{E}_{\boldsymbol{v}^\star,\sigma^\mu}\mathbb{E}_{\boldsymbol{\xi}^\mu|\boldsymbol{v}^\star,\sigma^\mu}[\,\cdot\,]$.

The $n-$replicated partition function, which was introduced in (6) to bypass the difficulty of performing the average of the log of Z), can be written as

$$Z^n \;=\; \int\prod_{\mu a}\frac{dh^{\mu a}d\hat{h}^{\mu a}}{2\pi}e^{i\sum_{\mu a}h^{\mu a}\hat{h}^{\mu a}-\beta\sum_{\mu a}\ell[\sigma^\mu(h^{\mu a}+b)]}\int\prod_{ia}dW_i^a\,e^{-\frac{\lambda}{2}\sum_{ia}(W_i^a)^2-\frac{i}{\sqrt{N}}\sum_{\mu a}\hat{h}^{\mu a}\sum_{i=1}^{N}W_i^a\xi_i^\mu}. \tag{A2}$$

where as usual we have assumed that n is integer. In the previous expression we have used a delta function and its integral representation in order to extract the scalar product $\boldsymbol{w}_i\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ in the loss argument. We can now perform the average over the pattern distribution, given the choice of the centroid and label; note that it is factorized over the number of examples in the training set. Indicating by $a,\,b\in\{1,\dots,n\}$ the replica indexes, we have

$$\begin{aligned}\prod_\mu\left[\mathbb{E}_{\boldsymbol{\xi}^\mu|\boldsymbol{v}^\star,\sigma^\mu}e^{-\frac{i}{\sqrt{N}}\sum_a\hat{h}^{\mu a}\sum_i w_i^a\xi_i^\mu}\right]&=\prod_\mu e^{-\frac{i}{N}\sum_a\sigma^\mu\hat{h}^{\mu a}\sum_i W_i^a v_i^*-\frac{\Delta}{2N}\sum_a\left(\hat{h}^{\mu a}\right)^2\sum_i(W_i^a)^2-\frac{\Delta}{N}\sum_{a<b}\hat{h}^{\mu a}\hat{h}^{\mu b}\sum_i W_i^aW_i^b}\\&=\int\prod_{a<b}\frac{dq^{ab}d\hat{q}^{ab}}{2\pi/N}\int\prod_a\frac{dQ^ad\hat{Q}^a}{2\pi/N}\frac{dM^ad\hat{M}^a}{2\pi/N}e^{-N\sum_{a<b}q^{ab}\hat{q}^{ab}+\sum_{a<b}\hat{q}^{ab}\sum_i W_i^aW_i^b}\\&\times e^{-N\sum_a Q^a\hat{Q}^a+\sum_a\hat{Q}^a\sum_i(W_i^a)^2-N\sum_a M^a\hat{M}^a+\sum_a\hat{M}^a\sum_i v_i^*W_i^a}\\&\times e^{-i\sum_{\mu a}\sigma^\mu\hat{h}^{\mu a}M^a-\frac{\Delta}{2}\sum_{\mu a}\left(\hat{h}^{\mu a}\right)^2Q^a-\Delta\sum_\mu\sum_{a<b}q^{ab}\hat{h}^{\mu a}\hat{h}^{\mu b}}.\end{aligned} \tag{A3}$$

In the last equality we have introduced several order parameters

- the overlap matrix between two weights $q^{ab}=\frac{1}{N}\sum_i w_i^aw_i^b$ for $a\neq b$;
- the squared norm $Q^a=\frac{1}{N}\sum_i(w_i^a)^2$;
- the overlap between a weight and the centroid $M^a=\frac{1}{N}\sum_i v_i^\star w_i^a$;
- the corresponding conjugated parameters that we denote by $\hat{q}^{ab}$, $\hat{Q}^a$ and $\hat{M}^a$ respectively that are used to enforce the previous definitions via Dirac delta functions.

The final expression for the replicated partition function is

$$\langle Z^n\rangle=\int\prod_{a<b}\frac{dq^{ab}d\hat{q}^{ab}}{2\pi/N}\int\prod_a\frac{dQ^ad\hat{Q}^a}{2\pi/N}\frac{dM^ad\hat{M}^a}{2\pi/N}e^{-N\sum_{a<b}q^{ab}\hat{q}^{ab}-N\sum_a Q^a\hat{Q}^a-N\sum_a M^a\hat{M}^a+NG_S+N\alpha G_E} \tag{A4}$$

where we have defined as usual an entropic and an energetic term

$$G_S\equiv\ln\mathbb{E}_{v^*}\int\prod_a dW^a\,e^{\sum_{a<b}\hat{q}^{ab}W^aW^b+\sum_a\hat{Q}^a(W^a)^2+v^*\sum_a\hat{M}^aW^a-\frac{\lambda}{2}\sum_a(W^a)^2} \tag{A5a}$$

$$G_E \equiv \ln \mathbb{E}_\sigma \int \prod_a \frac{dh^a d\hat{h}^a}{2\pi}\, e^{-\frac{\Delta}{2}\sum_a (\hat{h}^a)^2 Q^a - \Delta \sum_{a<b} q^{ab}\hat{h}^a\hat{h}^b + i\sum_a \hat{h}^a(h^a - \sigma M^a) - \beta \sum_a \ell[\sigma(h^a+b)]} \tag{A5b}$$

Note that in the entropic term the average over v^* can be done, so that we obtain

$$G_S = \ln \int \prod_a dW^a\, e^{\sum_{a<b}\left(\hat{q}^{ab} + \hat{M}^a\hat{M}^b\right)W^aW^b - \frac{1}{2}\sum_a \left(\lambda - \left(\hat{M}^a\right)^2 - 2\hat{Q}^a\right)(W^a)^2}\,. \tag{A6}$$

As usual in replica computations, we can now employ the saddle point method in order to evaluate the free entropy of the model. Correspondingly, the order parameters satisfy saddle point equations. Notice that learning the bias, that is having an additional integral over b in the definition of the partition function (5), corresponds to imposing additional saddle point equations for the replicated biases b^a, $a = 1, \dots, n$.

Appendix B: Typical case scenario: RS ansatz

In order to study typical solutions, we employ a replica-symmetric (RS) ansatz. This ansatz consists in seeking solutions to the saddle point equations of the form $q^{ab} = q$ for $a \neq b$, $Q^a = Q$, $M^a = M$ and similarly for conjugated order parameters. Defining $\mathcal{G}_S \equiv \lim_{n\to 0} G_S/n$ and $\mathcal{G}_E \equiv \lim_{n\to 0} G_E/n$, we have

$$\mathcal{G}_S = \frac{1}{2}\ln\left(\frac{2\pi}{\hat{q} - 2\hat{Q} + \lambda}\right) + \frac{\hat{q} + \hat{M}^2}{2\left(\hat{q} - 2\hat{Q} + \lambda\right)} \tag{B1a}$$

$$\mathcal{G}_E = \mathbb{E}_\sigma \int Dx \ln \int Dh\, e^{-\beta\ell\left(\sqrt{\Delta(Q-q)}h + \sqrt{\Delta q}x + M + \sigma b\right)} \tag{B1b}$$

where $Dx \equiv \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ is the standard Gaussian measure. The quenched free entropy $-\beta f \equiv \lim_{n\to 0} \frac{1}{nN}\ln\langle Z^n\rangle$ is simply

$$-\beta f = \frac{q\hat{q}}{2} - Q\hat{Q} - M\hat{M} + \mathcal{G}_S + \alpha\mathcal{G}_E \tag{B2}$$

All other quantities of interest can be computed from the free entropy. For example the training loss is

$$\epsilon_\ell = \frac{\partial(\beta f)}{\partial\beta} = \alpha\left\langle \ell\left(\sqrt{\Delta(Q-q)}h + \sqrt{\Delta q}x + M + \sigma b\right)\right\rangle_{x,h} \tag{B3}$$

where we have defined

$$\langle \mathcal{A}(x,h)\rangle_{x,h} \equiv \mathbb{E}_\sigma \int Dx\, \frac{\int Dh\, e^{-\beta\ell\left(\sqrt{\Delta(Q-q)}h + \sqrt{\Delta q}x + M + \sigma b\right)}\mathcal{A}(x,h)}{\int Dh\, e^{-\beta\ell\left(\sqrt{\Delta(Q-q)}h + \sqrt{\Delta q}x + M + \sigma b\right)}} \tag{B4}$$

The training error corresponding to a typical configuration of the loss ℓ is obtained simply by replacing the error counting function in (B3)

$$\epsilon_t = \alpha\mathbb{E}_\sigma\left\langle \Theta\left(-\sqrt{\Delta(Q-q)}h - \sqrt{\Delta q}x - M - \sigma b\right)\right\rangle_{x,h} \tag{B5}$$

where $\Theta(\cdot)$ is the Heaviside step function, that is $\Theta(x) = 1$ if $x \geq 0$ and 0 otherwise. Other important quantities such as the test loss and the test error are computed and reported in section E.

1. Error counting and MSE loss

Here we specialize the previous expressions in the case of error counting loss $\ell(x) = \Theta(-x)$ and the MSE $\ell(x) = \frac{1}{2}(x-1)^2$ one. In the error counting loss case the energetic term is

$$\mathcal{G}_E^{\text{err}} = \mathbb{E}_\sigma \int Dx \ln H_\beta\left(-\frac{\sqrt{\Delta q}x + M + \sigma b}{\sqrt{\Delta(Q-q)}}\right) \tag{B6}$$

where $H_\beta(x) \equiv e^{-\beta} + (1-e^{-\beta})H(x)$ and $H(x) \equiv \int_x^\infty Dh = \frac{1}{2}\text{erfc}\left(\frac{x}{\sqrt{2}}\right)$. Notice that in the error counting loss case the large β limit is trivial. In the MSE loss all the integrals in the energetic term can be performed, since they are all Gaussians. We obtain

$$\mathscr{G}_E^{\text{mse}} = -\frac{1}{2}\ln\left(1+\beta\Delta(Q-q)\right) - \frac{\beta}{2}\frac{\Delta q + (M-1)^2 + b^2 + 2b(2\rho-1)(M-1)}{1+\beta\Delta(Q-q)} \tag{B7}$$

2. Large β limit

We now concentrate on the case where the loss function has only one minimum. This happens for example in the previously mentioned case of the MSE loss. For those type of losses the overlap tends to the norm with a scaling given by

$$q = Q - \frac{\delta q}{\beta}\,. \tag{B8}$$

The conjugated parameters, as a consequence

$$\hat{q} = \beta^2\delta\hat{Q} - \beta\delta\hat{q} \tag{B9a}$$

$$\hat{Q} = \frac{\beta^2}{2}\delta\hat{Q} - \beta\delta\hat{q} \tag{B9b}$$

$$\hat{M} = \beta\delta\hat{M} \tag{B9c}$$

Notice also that in this limit the interesting regime of the regularization parameter is $\lambda \to \beta\lambda$. In this way the free energy in the large β limit is

$$-f = \mathfrak{G}_S + \alpha\mathfrak{G}_E \tag{B10}$$

where we have defined

$$\mathfrak{G}_S \equiv -\frac{\delta q\delta\hat{Q}}{2} + \frac{Q\delta\hat{q}}{2} - M\delta\hat{M} + \frac{\delta\hat{Q}+\delta\hat{M}^2}{2(\lambda+\delta\hat{q})} \tag{B11a}$$

$$\mathfrak{G}_E \equiv -\,\mathbb{E}_\sigma \int Dx\, A_\sigma(x) \tag{B11b}$$

and $A_\sigma(x) \equiv \min_h\left[\frac{h^2}{2} + \ell\left(\sqrt{\Delta\delta q}h + \sqrt{\Delta Q}x + M + b\sigma\right)\right]$. Calling by $h^*_\sigma(x)$ the corresponding argmin, the saddle point equations are

$$M = \frac{\delta\hat{M}}{\lambda+\delta\hat{q}} \tag{B12a}$$

$$Q = \frac{\delta\hat{Q}+\delta\hat{M}^2}{(\lambda+\delta\hat{q})^2} \tag{B12b}$$

$$\delta q = \frac{1}{\lambda+\delta\hat{q}} \tag{B12c}$$

$$\delta\hat{M} = \frac{\alpha}{\sqrt{\Delta\delta q}}\,\mathbb{E}_\sigma \int Dx\, h^*_\sigma(x) \tag{B12d}$$

$$\delta\hat{Q} = \frac{\alpha}{\delta q}\,\mathbb{E}_\sigma \int Dx\, \left[h^*_\sigma(x)\right]^2 \tag{B12e}$$

$$\delta\hat{q} = -\frac{\alpha}{\sqrt{Q\delta q}}\,\mathbb{E}_\sigma \int Dx\, x\, h^*_\sigma(x) = -\frac{\alpha}{\sqrt{Q\delta q}}\,\mathbb{E}_\sigma \int Dx\, \partial_x h^*_\sigma(x) \tag{B12f}$$

So far, we have not discussed how to fix the bias. One option is that considered in [19], where the bias b is learned in the same way of the weights w_i. This corresponds in adding an additional integral over b in the definition of the partition function (5). The whole replica analysis can be carried out in a similar way and it leads, in the RS ansatz, to an additional saddle point equation for the bias. This equation is

$$\mathbb{E}_\sigma\, \sigma \int Dx\, h^*_\sigma(x) = 0\,. \tag{B13}$$

Equations (B12) together with (B13) are the same reported in [19], with some change of variables. A second option is to fix the bias simply as an external parameter.

In the large β limit, the training loss becomes

$$\epsilon_\ell = \alpha\, \mathbb{E}_\sigma \int Dx\, \ell\left(\sqrt{\Delta\delta q}\, h^*_\sigma(x) + \sqrt{\Delta Q}x + M + \sigma b\right) ; \tag{B14}$$

The training error corresponding to a minimizer of the loss ℓ, is found, again, by plugging $\ell(x) = \Theta(-x)$ inside (B14). In the case of the MSE loss the energetic term simplifies as

$$\mathfrak{G}_E^{\rm mse} = -\frac{1}{2}\frac{\Delta Q + (M-1)^2 + b^2 + 2b(2\rho - 1)(M-1)}{1 + \beta\Delta\delta q} \tag{B15}$$

and the corresponding saddle point equations can be solved analytically as shown in [19]. For what follows it is important the expression of the training error, which is

$$\epsilon_t^{\rm mse} = \alpha\, \mathbb{E}_\sigma H\left(\frac{\Delta\delta q + M + b\sigma}{\sqrt{\Delta Q}}\right) , \tag{B16}$$

The training loss can be easily found by explicitly performing the y integral in (B14). One can verify that when λ is increased not only the corresponding squared norm Q lowers, but also, and more importantly, the training error/loss increases (even below the critical capacity $\alpha_c = 1$ of the model, where a zero training error solution can be found). This means that insisting in searching zero error solutions with the MSE loss is counterproductive and leads to overfitting. This is to be expected since the Gaussian mixture model is a particular case of general noisy teacher problems, in which the training set is no longer generated by a rule that can be inferred [28].

Appendix C: Measuring the local entropy of configurations using the Franz-Parisi approach

In order to compute the average over the disorder induced by the patterns, we use two replica tricks, one for the denominator of (7), which is just the partition function $\frac{1}{Z} = \lim_{r\to 0} Z^{r-1}$ and one for the log in the numerator of the same equation $\ln Z = \lim_{n\to 0} \partial_n Z^n$. From now on we will use indexes a or b for replicas in $\{1,\dots,r\}$ and c, $d \in \{1,\dots,n\}$. Therefore we get

$$\begin{aligned} -\beta f_{\rm FP}(S) &= \frac{1}{N}\lim_{n\to 0}\lim_{r\to 0}\partial_n \left\langle \int \prod_{a,i} d\tilde{w}_i^a \prod_a e^{-\beta_r \mathscr{L}_r(\tilde{\boldsymbol{w}}^a,\tilde{b}) - \frac{\lambda_r}{2}\sum_i (\tilde{w}_i^a)^2} \mathscr{V}^n(\tilde{\boldsymbol{w}}^{a=1}, S)\right\rangle \\ &= \frac{1}{N}\lim_{n\to 0}\lim_{r\to 0}\partial_n \langle Z_{\rm FP}^{n,r}\rangle \ . \end{aligned} \tag{C1}$$

The computation proceeds as usual by averaging over the disorder and introducing some other order parameters (in addition to those involving only the reference $\tilde{\boldsymbol{w}}$), namely $p^{cd} = \frac{1}{N}\sum_i w_i^c w_i^d$, $t^{ac} = \frac{1}{N}\sum_i \tilde{w}_i^a w_i^c$, $O^c = \frac{1}{N}\sum_i v_i^\star w_i^c$, $P^c = \frac{1}{N}\sum_i (w_i^c)^2$ and the corresponding conjugated ones. Note that P^c is just the squared norm P because of the spherical constraint inside the measure $d\mu_P(\boldsymbol{w})$[4]. Among the conjugated order parameters, we need also to introduce an additional parameter, $\hat{S}^c$, which is the one that impose the hard constraint on the overlap between the reference configuration $\boldsymbol{w}$ and $\boldsymbol{w}$. We finally obtain that the average of $Z_{\rm FP}^{n,r}$ is

$$\begin{aligned} \langle Z_{\rm FP}^{n,r}\rangle = &\int \prod_{a<b}\frac{dq^{ab}d\hat{q}^{ab}}{2\pi/N}\int\prod_{c<d}\frac{dp^{cd}d\hat{p}^{cd}}{2\pi/N}\int\prod_{a>1,c}\frac{dt^{ac}d\hat{t}^{ac}}{2\pi/N}\int\prod_a\frac{dQ^a d\hat{Q}^a}{2\pi/N}\frac{dM^a d\hat{M}^a}{2\pi/N} \\ &\int\prod_c \frac{d\hat{P}^c}{2\pi/N}\frac{dO^c d\hat{O}^c}{2\pi/N}\frac{d\hat{S}^c}{2\pi} e^{-N\sum_{a<b} q^{ab}\hat{q}^{ab} - N\sum_{c<d} p^{cd}\hat{p}^{cd} - N\sum_{a>1,c} t^{ac}\hat{t}^{ac}} \\ &\times e^{-N\sum_a Q^a\hat{Q}^a - NP\sum_c \hat{P}^c - N\sum_a M^a\hat{M}^a - N\sum_c O^c\hat{O}^c - NS\sum_c \hat{S}^c + NG_S + N\alpha G_E} \end{aligned} \tag{C2}$$

[4] We could have fixed the norm of the of the constrained configuration $\boldsymbol{w}$ by a Lagrange multiplier λ, in the same way of the reference. The two approaches (microcanonical versus canonical one) are however the same in the thermodynamic limit.

where

$$G_S \equiv \ln \int \prod_a d\tilde{W}^a \prod_c dW^c \, e^{\sum_{a<b}(\hat{q}^{ab}+\hat{M}^a\hat{M}^b)\tilde{W}^a\tilde{W}^b+\sum_{c<d}(\hat{p}^{cd}+\hat{O}^c\hat{O}^d)W^cW^d} \\ \times e^{\sum_{a>1,c}(\hat{t}^{ac}+\hat{M}^a\hat{O}^c)\tilde{W}^aW^c-\frac{1}{2}\sum_a(\lambda_r-(\hat{M}^a)^2-2\hat{Q}^a)(\tilde{W}^a)^2} \\ \times e^{\frac{1}{2}\sum_c((\hat{O}^c)^2+2\hat{P}^c)(W^c)^2+\tilde{W}^{a=1}\sum_c W^c(\hat{S}^c+\hat{M}^{a=1}\hat{O}^c)} \tag{C3}$$

and

$$G_E \equiv \ln \mathbb{E}_\sigma \int \prod_a \frac{dh^a d\hat{h}^a}{2\pi} \prod_c \frac{du^c d\hat{u}^c}{2\pi} e^{-\Delta\sum_{a<b} q^{ab}\hat{h}^a\hat{h}^b-\Delta\sum_{c<d}p^{cd}\hat{u}^c\hat{u}^d} \\ \times e^{-\Delta\sum_{a>1,c}t^{ac}\hat{h}^a\hat{u}^c-S\Delta\hat{h}^{a=1}\sum_c\hat{u}^c-\frac{\Delta}{2}\sum_a Q^a(\hat{h}^a)^2-\frac{\Delta}{2}P\sum_c(\hat{u}^c)^2} \\ \times e^{i\sum_a\hat{h}^a(h^a-\sigma M^a)+i\sum_c\hat{u}^c(u^c-\sigma O^c)-\beta_r\sum_a\ell_r\left[\sigma\left(h^a+\tilde{b}\right)\right]-\beta\sum_c\ell[\sigma(u^c+b)]} \, . \tag{C4}$$

1. RS ansatz

Next we impose, as usual, an RS ansatz, not only on the order parameters involving the reference as in appendix B, but also on those involving only $\boldsymbol{w}$ or mixing $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$, i.e. $p^{cd} = p$ for $c \neq d$, $t^{ac} = t$ for $a > 1$, $O^a = O$, $M^a = M$ and similarly for the conjugated ones. We finally get

$$-\beta f_{\rm FP}(S) = \frac{p\hat{p}}{2} + t\hat{t} - P\hat{P} - O\hat{O} - S\hat{S} + \mathscr{G}_S + \alpha\mathscr{G}_E \tag{C5}$$

where we have defined the entropic term as

$$\mathscr{G}_S = \frac{1}{\hat{p}-2\hat{P}}\left[\frac{(\hat{S}-\hat{t})^2\left(2(\hat{q}-\hat{Q})+\hat{M}^2+\lambda_r\right)}{2(\hat{q}-2\hat{Q}+\lambda_r)^2}+\frac{(\hat{S}-\hat{t})(\hat{t}+\hat{M}\hat{O})}{\hat{q}-2\hat{Q}+\lambda_r}\right]+\frac{\hat{p}+\hat{O}^2}{2(\hat{p}-2\hat{P})}+\frac{1}{2}\ln\left(\frac{2\pi}{\hat{p}-2\hat{P}}\right) . \tag{C6}$$

The energetic term instead is

$$\mathscr{G}_E = \mathbb{E}_\sigma \int Dx \, \frac{1}{\mathscr{Z}(x)} \int Dh \, e^{-\beta_r\ell_r\left(\sigma\tilde{b}+M+\sqrt{\Delta q}x+\sqrt{\Delta(Q-q)}h\right)} \\ \times \int Dy \ln \int Du \, e^{-\beta\ell\left[\sigma b+O+\sqrt{\Delta\gamma-\frac{\Delta(S-t)^2}{Q-q}}y+\frac{\Delta t}{\sqrt{\Delta q}}x+\frac{\Delta(S-t)}{\sqrt{\Delta(Q-q)}}h+\sqrt{\Delta(P-p)}u\right]} . \tag{C7}$$

In the previous equation have defined $\gamma = p - \frac{t^2}{q}$ and

$$\mathscr{Z}(x) \equiv \int Dh \, e^{-\beta_r\ell_r\left(\sigma\tilde{b}+M+\sqrt{\Delta q}x+\sqrt{\Delta(Q-q)}h\right)} . \tag{C8}$$

Note that the parameters involving only the reference $\tilde{\boldsymbol{w}}$, i.e. q, $\hat{q}$, $\hat{Q}$, M and $\hat{M}$ satisfy the same saddle point equations of the previous subsection [25].

2. Low temperature limit for the reference configuration

We are now interested in sending β_r to infinity. In order to do that, we need to add to the scalings of the order parameters involving only the reference (B9), together with the ones for the overlaps between reference $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ and their conjugated ones. These are readily seen to be

$$t = S - \frac{\delta t}{\beta_r} \tag{C9a}$$

$$\hat{t} = \beta_r \, \delta\hat{t} \tag{C9b}$$

$$\hat{S} - \hat{t} = \delta\hat{S}\,. \tag{C9c}$$

Using these scalings, the Franz-Parisi entropy is given by equation (10). We have accordingly redefined the entropic and energetic terms, respectively as

$$\begin{aligned}\mathfrak{G}_S = &\frac{p\hat{p}}{2} - \delta t\hat{\delta}t - P\hat{P} - O\hat{O} - S\delta\hat{S} + \frac{1}{2}\ln\left(\frac{2\pi}{\hat{p}-2\hat{P}}\right)\\ &+ \frac{1}{\hat{p}-2\hat{P}}\left[\frac{\hat{p}+\hat{O}^2}{2} + \frac{\delta\hat{S}^2\left(\delta\hat{Q}+\delta\hat{M}^2\right)}{2(\delta\hat{q}+\lambda_r)^2} + \frac{\delta\hat{S}(\delta\hat{t}+\delta\hat{M}\hat{O})}{\delta\hat{q}+\lambda_r}\right],\end{aligned} \tag{C10a}$$

$$\mathfrak{G}_E = \mathbb{E}_\sigma \int Dx Dy \ln \int Du\, e^{-\beta\ell\left[\sigma b + O + \sqrt{\Delta\Gamma}y + \frac{\Delta S}{\sqrt{\Delta Q}}x + \frac{\Delta\delta t}{\sqrt{\Delta\delta q}}h^*_\sigma(x) + \sqrt{\Delta(P-p)}u\right]}. \tag{C10b}$$

In the last expression we have defined $\Gamma = p - \frac{S^2}{Q}$.

Once $f_{\rm FP}$ is known (by solving the corresponding saddle point equations), we can compute the local energy ϵ_ℓ as $\epsilon_\ell = \frac{\partial(\beta f_{\rm FP})}{\partial\beta}$ and the local entropy as $\mathcal{S} = \beta(\epsilon_\ell - f_{\rm FP})$. The same formulas are valid if we look to the local entropy landscape in the space of the *training error*, where $\ell(x) = \Theta(-x)$. For clarity we denote the training error as ϵ_t, as in the previous subsection. The training error can be computed by

$$\epsilon_t = \alpha e^{-\beta}\mathbb{E}_\sigma \int Dx Dy \frac{H\left(\frac{\sigma b + O + \sqrt{\Delta\Gamma}y + \frac{\Delta S}{\sqrt{\Delta Q}}x + \frac{\Delta\delta t}{\sqrt{\Delta\delta q}}h^*_\sigma(x)}{\sqrt{\Delta(P-p)}}\right)}{H_\beta\left(-\frac{\sigma b + O + \sqrt{\Delta\Gamma}y + \frac{\Delta S}{\sqrt{\Delta Q}}x + \frac{\Delta\delta t}{\sqrt{\Delta\delta q}}h^*_\sigma(x)}{\sqrt{\Delta(P-p)}}\right)} \tag{C11}$$

where $H_\beta(x) \equiv e^{-\beta} + (1-e^{-\beta})H(x)$.

Appendix D: Large deviation case: 1RSB ansatz

We now study a system of y real replicas where each replica optimizes a loss ℓ under constraints on their squared norm Q and also on their mutual angles, namely: $\forall a, b:\ \frac{1}{N}\sum_i (w_i^a)^2 = Q,\ \frac{1}{N}\sum_i w_i^a w_i^b = Q\cos(\theta)$. Thus the partition function of this system of y real replicas is

$$Z_y = \int \prod_a d\mu_Q(\boldsymbol{w}^a)\, e^{-\beta\sum_a \mathscr{L}(\boldsymbol{w}^a, b)}\,\delta\left(Q\cos(\theta) - \frac{1}{N}\sum_i w_i^a w_i^b\right) \tag{D1}$$

Notice that here we are imposing the constraint over the squared norm via an hard constraint (instead of a soft constraint as was used in appendix B). Note, in addition, that the constraint is imposed only on the weights; the biases are assumed to be all equal in all the replicas. The free entropy of a single replica is $-\beta f = \frac{1}{Ny}\overline{\ln Z_y}$ and can be evaluated by the usual replica trick

$$-\beta f = \lim_{s\to 0}\frac{1}{Ny}\partial_s \overline{Z^s} \tag{D2}$$

It is now evident that if we choose $s = n/y$ virtual replicas, studying a system of y real replicas constrained to be at a mutual overlap q_1 is equivalent to impose a 1RSB ansatz on the standard equilibrium measure (5) with the Parisi parameter m and the intra-block overlap parameter q_1 fixed as external parameters; they play the same role respectively of the number of replicas y and the overlap between replicas $Q^2\cos(\theta)$ (see also [7, 29] for a detailed analysis). The final result is

$$-\beta f = \frac{q_1\hat{q}_1}{2} - \frac{y}{2}(q_1\hat{q}_1 - q_0\hat{q}_0) - Q\hat{Q} - M\hat{M} + \mathcal{G}_S + \alpha\mathcal{G}_E \tag{D3}$$

where

$$\mathcal{G}_S = \frac{1}{2}\ln\left(\frac{2\pi}{\hat{q}_1 - 2\hat{Q}}\right) + \frac{1}{2}\frac{\hat{q}_0 + \hat{M}^2}{\hat{q}_1 - 2\hat{Q} - y(\hat{q}_1 - \hat{q}_0)} + \frac{1}{2y}\ln\left(\frac{\hat{q}_1 - 2\hat{Q}}{\hat{q}_1 - 2\hat{Q} - y(\hat{q}_1 - \hat{q}_0)}\right) \tag{D4a}$$

$$\mathscr{G}_E = \frac{1}{y}\mathbb{E}_\sigma \int Dx \, \ln \int Dz \left[\int Dh \, e^{-\beta \ell\left(\sqrt{\Delta(Q-q_1)}h + \sqrt{\Delta q_0}x + \sqrt{\Delta(q_1-q_0)}z + M + \sigma b\right)} \right]^y . \tag{D4b}$$

Notice that in the $q_1 \to Q$ limit, the expressions reduce to the RS case given in eq. (B1b), but with β replaced by βy. Again, in the particular case of the MSE loss, all the integrals in the energetic term can be solved, giving

$$\begin{aligned}\mathscr{G}_E^{\text{mse}} = &-\frac{1}{2}\ln\left(1+\beta\Delta(Q-q_1)\right) + \frac{1}{2y}\ln\left(\frac{1+\beta\Delta(Q-q_1)}{1+\beta\Delta(Q-q_1)+y\beta\Delta(q_1-q_0)}\right) \\ &- \frac{\beta}{2}\frac{\Delta q_0 + \mathbb{E}_\sigma(M+b\sigma-1)^2}{1+\beta\Delta(Q-q_1)+\beta y\Delta(q_1-q_0)} .\end{aligned} \tag{D5}$$

1. Computing the barycenter of the replicas

From the system of y real replicas one can compute new valid configurations of the weights that can achieve good generalization properties. As shown in [3] one of those is the barycenter of the replicated system, defined as

$$\overline{w}_i \equiv \frac{1}{y}\sum_a w_i^a . \tag{D6}$$

The barycenter is identified by its overlap with the teacher $\overline{M} = \frac{1}{N}\sum_i \overline{w}_i v_i^\star$ and its squared norm $\overline{Q} = \frac{1}{N}\sum_i \overline{w}_i^2$. They can computed by expressing them in terms of the known replica-overlap quantities. In fact

$$\overline{M} = \frac{1}{Ny}\sum_i\sum_a w_i^a v_i^\star = \frac{1}{y}\sum_a M^a \tag{D7a}$$

$$\overline{Q} = \frac{1}{N}\sum_i \overline{w}_i^2 = \frac{1}{y^2 N}\sum_{ab}\sum_i w_i^a w_i^b = \frac{1}{y^2}\sum_a Q^a + \frac{1}{y^2}\sum_{a\neq b} q^{ab} \tag{D7b}$$

Since all real replicas are have the same mutual overlap and the same squared norm, we get

$$\overline{M} = M \tag{D8a}$$

$$\overline{Q} = \frac{Q-q_1}{y} + q_1 \tag{D8b}$$

Notice that since $|q_1| < Q$, the squared norm of the center is always lower than that of a real replica. In addition, there is always a maximal angle at which the replicas can be placed. This value is

$$\theta_{\max} = \arccos\left(-\frac{1}{y-1}\right) \tag{D9}$$

for example, when $y = 3$, the maximal angle is clearly $\theta_{\max} = \frac{2\pi}{3}$, i.e. the replicas are placed on the vertices of a equilateral triangle. In general at the maximal angle the y replicas are placed on the vertices of a regular y-simplex[5] immersed on the $(N-1)$-sphere of radius $\sqrt{N}$.

2. Large-β limit for the MSE loss

Let us now perform the large β limit in the case of a loss having only one minimum. We restrict for simplicity to the MSE loss case.

It can be seen that the correct scaling of q_0 is

$$q_0 = \frac{Q-q_1}{y} + q_1 - \frac{\delta q_0}{\beta} . \tag{D10}$$

---

[5] Since in an y-simplex $\sum_{a=1}^{y} \boldsymbol{u}^a = 0$ where $||\boldsymbol{u}^a|| = 1$ are unit-norm vectors that identify the position of the vertices with respect to the barycenter, we have $||\sum_{a=1}^{y} \boldsymbol{u}^a||^2 = y + y(y-1)\cos(\theta) = 0$ which gives eq. (D9).

and the conjugated parameters instead scale as

$$\hat{q}_0 = \beta^2 \delta\hat{q}_0 + \frac{\beta}{2}\delta\hat{q}_1 \tag{D11a}$$

$$\hat{q}_1 = \beta^2 \delta\hat{q}_0 - \frac{\beta}{2}\delta\hat{q}_1 \tag{D11b}$$

$$\hat{Q} = \frac{\hat{q}_1}{2} - \frac{1}{2(Q-q_1)} \tag{D11c}$$

$$\hat{M} = \beta\delta\hat{M} \tag{D11d}$$

The free energy is therefore

$$-f = \frac{1}{2}(Q-q_1)\delta\hat{q}_1 + \frac{y}{2}\left[q_1\delta\hat{q}_1 - \delta q_0\delta\hat{q}_0\right] - M\delta\hat{M} + \mathfrak{G}_S + \alpha\mathfrak{G}_E \tag{D12}$$

where the new entropic and energetic terms (rescaled with β) are

$$\mathfrak{G}_S^{\rm mse} \equiv \lim_{\beta\to\infty} \frac{\mathcal{G}_S^{\rm mse}}{\beta} = \frac{1}{2}\,\frac{\delta\hat{q}_0 + \delta\hat{M}^2}{y\delta\hat{q}_1} \tag{D13a}$$

$$\mathfrak{G}_E^{\rm mse} \equiv \lim_{\beta\to\infty} \frac{\mathcal{G}_E^{\rm mse}}{\beta} = -\frac{1}{2}\frac{\Delta\left(\frac{Q-q_1}{y}+q_1\right) + \mathbb{E}_\sigma(M+b\sigma-1)^2}{1+y\Delta\delta q_0} \tag{D13b}$$

The training error is

$$\epsilon_t^{\rm mse} = \alpha\mathbb{E}_\sigma H\left(\frac{\Delta\delta q_0 + M + b\sigma}{\Delta\left(\frac{Q-q_1}{y}+q_1\right)}\right) \tag{D14}$$

Notice that in the large β limit the training error/loss of one of the replicas is not zero, because of distance constraint and the convexity of the loss landscape.

The large y-limit can also be handled (in a way similar to the case of a generic loss): it suffices to rescale the order parameters $\delta q_0 \to {}^{\delta q_0}/_{y}$ and $\delta\hat{q}_1 \to {}^{\delta\hat{q}_1}/_{y}$.

Appendix E: Test loss and generalization error

Let us compute the test loss of a configuration of weights $\boldsymbol{w}$ that has overlap with the teacher M and squared norm Q after a training with αN patterns. In order to do that, we need to present to $\boldsymbol{w}$ a new example that we will denote by $\boldsymbol{\xi}$ with its corresponding label σ. The test loss is therefore

$$\mathscr{L}_t(\boldsymbol{w},\boldsymbol{v}^\star) \equiv \mathbb{E}_\sigma\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{v}^\star,\sigma}\left\langle \ell\left[\sigma\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N w_i\,\xi_i + b\right)\right]\right\rangle_Z \tag{E1}$$

where $\langle\bullet\rangle_Z$ is the average over $\boldsymbol{w}$ extracted according to the partition function given in (5). Extracting the loss argument by a delta function and performing the expectation over the pattern distribution, we finally get

$$\mathscr{L}_t(\boldsymbol{w},\boldsymbol{v}^\star) = \mathbb{E}_\sigma \int Du\,\ell\left(\sigma b + M + \sqrt{\Delta Q}u\right). \tag{E2}$$

In the case of the error counting loss $\ell(x) = \Theta(-x)$ we get the expression for the generalization error given in [19]

$$\epsilon_g = \rho H\left(\frac{M+b}{\sqrt{\Delta Q}}\right) + (1-\rho)H\left(\frac{M-b}{\sqrt{\Delta Q}}\right). \tag{E3}$$

For the MSE loss instead we obtain

$$\mathscr{L}_t^{\rm mse}(\boldsymbol{w},\boldsymbol{v}^\star) = \frac{1}{2}\left[\Delta Q + \mathbb{E}_\sigma(M-1+\sigma b)^2\right]. \tag{E4}$$

Appendix F: Bayesian generalization error

In this section we derive the expression of the Bayesian generalization error. For a given generator $\boldsymbol{v}^\star$ and a configuration of the weights $\boldsymbol{w}$ the generalization error is found by extracting a test pattern $\boldsymbol{\xi}$ with label σ and computing the corresponding error

$$\begin{aligned}\epsilon_g(\boldsymbol{w},\boldsymbol{v}^\star) &= \mathbb{E}_\sigma \int d\boldsymbol{\xi}\, P(\boldsymbol{\xi},\sigma|\boldsymbol{v}^\star)\Theta\left(-\sigma\hat{\sigma}(\boldsymbol{\xi};\boldsymbol{w},b)\right)\\ &= \mathbb{E}_\sigma \int Du\,\Theta\left(-\sigma b - \frac{1}{N}\sum_i w_i v_i^\star - \sqrt{\frac{\Delta}{N}\sum_i w_i^2}\,u\right)\\ &= \mathbb{E}_\sigma H\left(\frac{\sigma b + \frac{1}{N}\sum_i w_i v_i^\star}{\sqrt{\frac{\Delta}{N}\sum_i w_i^2}}\right). \end{aligned} \tag{F1}$$

The Bayesian mean generalization error is obtained by averaging $\epsilon_g(\boldsymbol{w},\boldsymbol{v}^\star)$ with respect to the posterior $P(\boldsymbol{v}|\,\{\boldsymbol{\xi}^\mu,\sigma^\mu\})$. Applying Bayes theorem we have

$$\begin{aligned}\epsilon_g^B(\boldsymbol{w}|\,\{\boldsymbol{\xi}^\mu,\sigma^\mu\}) &= \int d\boldsymbol{v}^\star\, P(\boldsymbol{v}^\star|\,\{\boldsymbol{\xi}^\mu,\sigma^\mu\})\epsilon_g(\boldsymbol{w},\boldsymbol{v}^\star) = \frac{\int d\boldsymbol{v}^\star\, P(\{\boldsymbol{\xi}^\mu,\sigma^\mu\}\,|\boldsymbol{v}^\star)P(\boldsymbol{v}^\star)\,\epsilon_g(\boldsymbol{w},\boldsymbol{v}^\star)}{\int d\boldsymbol{v}^\star\, P(\{\boldsymbol{\xi}^\mu,\sigma^\mu\}\,|\boldsymbol{v}^\star)P(\boldsymbol{v}^\star)}\\ &= \mathbb{E}_\sigma H\left(\frac{\sigma b + \frac{1}{N}\sum_i w_i \overline{v}_i}{\sqrt{\frac{\Delta}{N}\sum_i w_i^2}}\right), \end{aligned} \tag{F2}$$

where we have defined

$$\overline{v}_i \equiv \frac{1}{\Delta+\alpha}\frac{\sum_\mu \xi_i^\mu \sigma^\mu}{\sqrt{N}}. \tag{F3}$$

Minimizing (F2) with respect to $\boldsymbol{w}$ and b, we obtain

$$\boldsymbol{w}^{\rm opt}(\{\boldsymbol{\xi}^\mu,\sigma^\mu\}) = \overline{\boldsymbol{v}} \tag{F4a}$$

$$b^{\rm opt} = \frac{\Delta}{2}\ln\left(\frac{\rho}{1-\rho}\right). \tag{F4b}$$

We can now use $\boldsymbol{w}^{\rm opt}$ and $b^{\rm opt}$ to compute the minimal expected generalization error: we first extract a $\boldsymbol{v}^\star$, with that a training set $\{\boldsymbol{\xi}^\mu,\sigma^\mu\}$, from that compute $\boldsymbol{w}^{\rm opt}$ and $b^{\rm opt}$, then extract a test pattern $\boldsymbol{\xi}^\star$ with label $\sigma^\star$ and finally use it to compute the error. We have

$$\epsilon_g^B = \int d\boldsymbol{v}^\star\, d\boldsymbol{\xi}^\star\, d\sigma^\star\, P(\boldsymbol{\xi}^\star,\sigma^\star|\boldsymbol{v}^\star)\int\prod_\mu d\boldsymbol{\xi}^\mu d\sigma^\mu \prod_\mu P(\boldsymbol{\xi}^\mu,\sigma^\mu|\boldsymbol{v}^\star)P(\boldsymbol{v}^\star)\,\Theta\left(-\sigma^\star\hat{\sigma}(\boldsymbol{\xi}^\star;\boldsymbol{w}^{\rm opt},b^{\rm opt})\right). \tag{F5}$$

By using the central limit theorem repeatedly, we obtain

$$\epsilon_g^B = \mathbb{E}_{\sigma^\star} H\left(\frac{M^{\rm opt} + b^{\rm opt}\sigma^\star}{\sqrt{\Delta Q^{\rm opt}}}\right), \tag{F6}$$

where we have defined

$$M^{\rm opt} = Q^{\rm opt} \equiv \frac{\alpha}{\Delta+\alpha}. \tag{F7}$$

Appendix G: Numerical details

We used rSGD for all simulations reported in this work. As described in the main text, the algorithm consists in training y replicas of a perceptron each initialized in a different way, with an additional term in the loss function of each model proportional to the sum of distances from the other replicas. The total loss function is

$$\begin{aligned}\mathscr{L}(\{\boldsymbol{w}^a, b^a\}_{a=1}^y) &= \sum_{a=1}^{y} [\mathscr{L}^a_{\text{MSE}} + \lambda \mathscr{L}^a_{\text{d}}] \\ &= \sum_{a=1}^{y} \left[\mathscr{L}^a_{\text{MSE}} + \lambda \sum_{a\neq b}^{y} (d_{ab} - d_0)^2\right]\end{aligned} \tag{G1}$$

where $\mathscr{L}^a_{\text{d}}$ is the term we introduced in order to force the replicas to stay at a given distance d_0. The bias is treated separately: in the cases where $\rho = 0.5$ it is simply set to zero; in the cases where $\rho \neq 0.5$ we add to the loss of each replica $\lambda \sum_{a\neq b}^{y} (b_a - b_b)^2$. This is done in order to match results with analytical calculations, where the replicas share the same bias.

Then we define a center model as our predictor, defined as the average of the replicas in the following way:

$$\bar{\boldsymbol{w}} = \frac{1}{y} \sum_{a=1}^{y} \boldsymbol{w}^a \tag{G2a}$$

$$\bar{b} = \|\bar{\boldsymbol{w}}\| \frac{1}{y} \sum_{a=1}^{y} \frac{b^a}{\|\boldsymbol{w}^a\|} \tag{G2b}$$

The perceptron $(\bar{\boldsymbol{w}}, \bar{b})$ is the model we use to compute loss and error on both the testset and the trainset. Note that, as discussed in the main text, the bias of the center model is not simply the average of the biases, but rather the average of the biases weighted by the inverse norm of the weights, scaled by the norm of the predictor itself. Note that at the end of the training the replicas are expected to have the same value b of the bias, and since the norm of the replicas is fixed to some given $\|\boldsymbol{w}\|$ the bias of the center will simply be $\bar{b} = b\,\|\bar{\boldsymbol{w}}\| \,/\, \|\boldsymbol{w}\|$.

Most of the following details should not matter for the sake of generalization error because the problem is convex. They still determine the rate of convergence to the analytical solution, so we report them in detail.

The training is performed with PyTorch by using full-batch gradient descent with learning rate $1{\cdot}10^{-4}$. Initialization is standard Xavier. In the cases with $y = 1$ we train with the Adam optimizer for $2 \cdot 10^4$ epochs.

In the cases with $y > 1$ we train with the SGD optimizer with momentum 0.5 for $4 \cdot 10^4$ epochs. In those cases we increase the coupling constant λ at each epoch by a factor $\lambda_1 = 5 \cdot 10^{-3}$ starting from the value $\lambda_0 = 1 \cdot 10^{-4}$ up to a maximum $\lambda_{\max} = 1 \cdot 10^2$; namely we set $\lambda(t) = \min[\lambda_0(1 + \lambda_1)^t, \lambda_{\max}]$.

The norm is always kept fixed by renormalizing the weights to the given magnitude before each forward pass of the perceptron.

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms* (Cambridge university press, 2014).

[2] S. Hochreiter and J. Schmidhuber, Neural Computation **9**, 1 (1997), https://doi.org/10.1162/neco.1997.9.1.1.

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

[4] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, CoRR **abs/1609.04836** (2016), arXiv:1609.04836.

[5] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," (2019), arXiv:1912.02178 [cs.LG].

[6] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," (2017), arXiv:1703.11008 [cs.LG].

[7] C. Baldassi, F. Pittorino, and R. Zecchina, Proceedings of the National Academy of Sciences **117**, 161 (2020), https://www.pnas.org/content/117/1/161.full.pdf.

[8] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Proceedings of the National Academy of Sciences **113**, E7655 (2016), https://www.pnas.org/content/113/48/E7655.full.pdf.

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

[10] F. Borra, M. C. Lagomarsino, P. Rotondo, and M. Gherardi, Journal of Physics A: Mathematical and Theoretical **52**, 384004 (2019).
[11] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Phys. Rev. E **102**, 032119 (2020).
[12] P. Rotondo, M. Pastore, and M. Gherardi, Phys. Rev. Lett. **125**, 120601 (2020).
[13] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, (2019).
[14] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, arXiv preprint arXiv:2002.09339 (2020).
[15] X. Mai and Z. Liao, arXiv preprint arXiv:1905.13742 (2019).
[16] M. Lelarge and L. Miolane, arXiv preprint arXiv:1907.03792 (2019).
[17] Z. Deng, A. Kammoun, and C. Thrampoulidis, arXiv preprint arXiv:1911.05822 (2019).
[18] T. Lesieur, C. De Bacco, J. Banks, F. Krzakala, C. Moore, and L. Zdeborová, in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (IEEE, 2016) pp. 601–608.
[19] F. Mignacco, F. Krzakala, Y. M. Lu, and L. Zdeborová, arXiv preprint arXiv:2002.11544 (2020).
[20] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Journal of Statistical Mechanics: Theory and Experiment **2016**, P023301 (2016).
[21] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. T. Chayes, L. Sagun, and R. Zecchina, in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (OpenReview.net, 2017).
[22] M. Welling and Y. W. Teh, in *Proceedings of the 28th international conference on machine learning (ICML-11)* (2011) pp. 681–688.
[23] F. Pittorino, C. Lucibello, C. Feinauer, E. M. Malatesta, G. Perugini, C. Baldassi, M. Negri, E. Demyanenko, and R. Zecchina, arXiv preprint arXiv:2006.07897 (2020).
[24] S. Franz and G. Parisi, Journal de Physique I **5**, 1401 (1995).
[25] H. Huang and Y. Kabashima, Physical Review E **90**, 052813 (2014).
[26] E. Gardner, Journal of Physics A: Mathematical and General **21**, 257 (1988).
[27] E. Gardner and B. Derrida, Journal of Physics A: Mathematical and General **21**, 271 (1988).
[28] A. Engel and C. Van den Broeck, *Statistical mechanics of learning* (Cambridge University Press, 2001).
[29] R. Monasson, Physical review letters **75**, 2847 (1995).