A FLATNESS-GENERALIZATION RELATION IN THE TEACHER-STUDENT TREE-COMMITTEE MACHINE

Anonymous authors
Paper under double-blind review

ABSTRACT

The flatness of the loss landscape at a minimizer is a widely used heuristic for reasoning about neural-network generalization, yet evidence for this relation is mostly empirical and controversial. We study this relation in a teacher-student tree committee machine, where both the ERM estimator and the Hessian spectrum are analytically tractable in the proportional high-dimensional limit. First, we use a zero-temperature Gibbs formulation to obtain predictions for the observables of the typical minimizers of the empirical loss. Secondly, we use Edwards-Jones formalism to derive the limiting Hessian resolvent around these typical minimizers. All predictions agree with finite-size gradient-descent simulations. Finally, we study three measures of flatness, namely the left and right edges and the spectral mean, and check if a decrease in generalization error as the dataset size is increased corresponds to an increase in flatness. We find that the answer strongly depends on the learning task and on the ratio of the number of parameters to the number of data points. In regression, the spectral mean and right edge correlate with the generalization error, while the left edge does so only in the overparametrized regime. In classification this correlation reliably holds only in the highly overparametrized phase, while for underparametrized networks it can even reverse.

1 INTRODUCTION

Why do overparametrized neural networks generalize after fitting high-dimensional training data? Among the many theoretical approaches that theoreticians have proposed to answer this question, a classic one has been to study the geometry of the loss landscape, in particular relating the generalization capacity of a network to the flatness of the minimizer (Hochreiter & Schmidhuber, 1997; MacKay, 1992; Hinton & van Camp, 1993; Hochreiter & Schmidhuber, 1994). The simple and heuristic argument is that if the difference between the population and empirical loss is not too big, then a function that minimizes the empirical loss and is also flat should also perform well on the population. This argument has been supported by many empirical works. For example (Keskar et al., 2016) show that increasing the batch size of SGD leads to both worse generalization and sharper minima. Furthermore, the notion of flatness has been used to propose new optimization algorithms that favor flat solutions, which typically perform better in terms of generalization (Foret et al., 2021; Chaudhari et al., 2019).

Although this flatness-generalization relation is deeply rooted in machine learning lore, many subsequent works have found settings where this relation is broken (Zhang et al., 2021; Shoham et al., 2025; Andriushchenko et al., 2023). For example (Dinh et al., 2017) show that fully connected networks with ReLU activations can be reparametrized in a way that keeps the output invariant but varies many Hessian-based measures of flatness, thus potentially breaking this relation for any such measure. In an attempt to restore the relation, many works have come up with alternative measures of flatness that are invariant to this rescaling (Tsuzuku et al., 2020; Shoham et al., 2025; Petzka et al., 2020; Pittorino et al., 2022; Kwon et al., 2021). Another line of research has compared the flatness of all minimizers of a loss, showing in some cases that minimizers with bad generalization properties can never be the flattest, even under reparametrization (Vardhan et al., 2026; Ding et al., 2024).

A widely used tool for quantifying local flatness is the Hessian of the empirical loss. Its largest eigenvalue measures the curvature along the sharpest direction and arises from a quadratic approximation to worst-case neighborhood-based sharpness, whereas its trace measures the average

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

curvature over parameter-space directions. Early empirical investigations found that the Hessian spectra of trained neural networks typically consist of a highly degenerate bulk concentrated near zero together with a small number of isolated eigenvalues (Sagun et al., 2017; Baldassi et al., 2020). Subsequent work developed scalable methods for estimating the entire spectrum and studied its evolution during training, revealing that the gradient can become strongly aligned with the eigenspaces associated with the largest eigenvalues (Ghorbani et al., 2019). Other studies connected these spectral structures to batch size and robustness (Yao et al., 2018), and showed that outliers and secondary bulks can encode class and cross-class structure in the data (Papyan, 2020; Sabanayagam et al., 2023). A unified phenomenological model proposed by (Fort & Ganguli, 2019) connected several of these observations: the emergence of a small number of directions with large positive curvature, the alignment of the gradient with the corresponding eigenspace, and the non-monotonic evolution of the largest Hessian eigenvalue during training.

On the theoretical side, random-matrix methods have been used to characterize the Hessian spectrum of simple neural network models (Pennington & Bahri, 2017; Liao & Mahoney, 2021; Franz et al., 2015; Asgari et al., 2025; Montanari & Saeed, 2026). In the case of networks with quadratic activations, related analyses have shown that spectral transitions of the Hessian can mark changes in the stability or informativeness of stationary points (Sarao Mannelli et al., 2020; Bonnaire et al., 2024; Annesi et al., 2025b). However, in none of these works has the Hessian been studied from a flatness-generalization perspective.

The advantage of these models is that they offer a controlled setting in which macroscopic observables can be characterized analytically. For example, in teacher-student settings, data is labeled by a teacher network and used to train a student, so that generalization is determined by a small number of teacher–student and student–student overlaps. Statistical-mechanics methods have long been used in this setting to derive learning curves, phase transitions, and asymptotically exact dynamical equations (Watkin et al., 1993; Engel & Van den Broeck, 2001; Biehl & Schwarze, 1995; Saad & Solla, 1995). Among the architectures studied in this framework, committee machines are particularly useful because they combine analytical tractability with important features of more complex models, including non-convexity and hidden-unit specialization. Both classical and modern works have used the replica method to uncover interesting learning phenomena such as SAT/UNSAT transitions (Barkai et al., 1992; 1990; Engel et al., 1992; Annesi et al., 2025a; Zavatone-Veth & Pehlevan, 2021; Baldassi et al., 2019), specialization (Schwarze & Hertz, 1992; Biehl et al., 1998; Ahr et al., 1999; Barbier et al., 2025), and statistical-to-computational gaps (Aubin et al., 2018).

Contributions. In this work, we investigate the flatness–generalization relation in teacher–student tree committee machines in the proportional high-dimensional limit for two different learning tasks, regression and classification. We first use the replica method to characterize typical minimizers of the loss and derive analytical predictions for the order parameters that describe them. Depending on the learning task, the dataset size and dimensionality, and the regularization strength, we find that the solution is either replica symmetric (RS) or exhibits full replica symmetry breaking (FRSB). Intuitively, this means that in the first case the set of minimizers is connected, while in the latter, the low-loss configurations are organized into a continuous hierarchy of clusters characterized by different mutual overlaps, as in Parisi's solution of mean-field spin glasses (Mézard et al., 1987). Although we are able to solve the FRSB equations for a set of parameters of the model, we then approximate it with a 1RSB solution, which is numerically easier to handle, and which we confirm to be a good approximation of FRSB observables (Annesi et al., 2025a).

Secondly, we use the Edwards-Jones replica formalism (Edwards & Jones, 1976) to derive the spectrum of the Hessian matrix calculated on these minimizers. For square loss the resolvent is obtained from a scalar cubic equation whose coefficients are functions of the same order parameters that describe learning. From this, we calculate a set of scalar observables which are commonly used in the literature to quantify the flatness of a minimum, mainly the edges of the spectral distribution and its mean. Then, by varying dataset size, we check if a flatness-generalization correlation holds for some of these observables.

We find that whether this relation holds or not is highly dependent on the observable, the learning task, and the parameters. For regression, we find that the relation holds both for the right edge (known in the literature as sharpness) and for the trace, while it holds for the left edge only in the overparametrized phase. For classification instead, we find that it holds for all three observables only

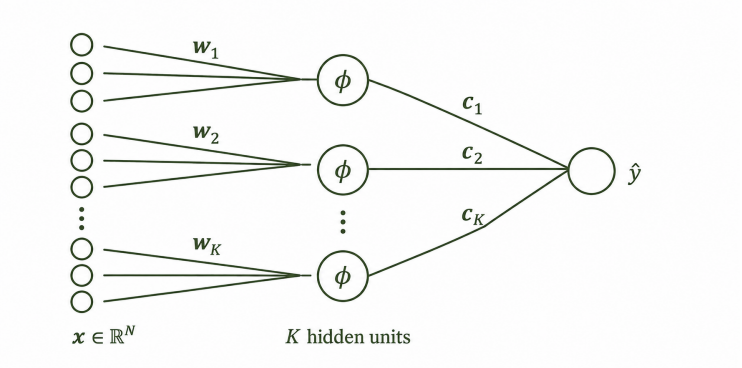

Figure 1: Schematic representation of a tree committee machine.

in the overparametrized phase: in all other phases the relation is broken, and in some cases flatness can also anticorrelate with generalization.

Finally, we validate all our theoretical predictions with an extensive set of gradient descent simulations. We find good agreement with the theory even in the replica symmetry broken regime, where the exact equilibrium sampling is known to be challenging for stable algorithms in related mean field spin glass models El Alaoui et al. (2022).

2 The Model

We consider a teacher–student learning problem where both the student and teacher are tree-committee machines. This is a two-layer neural network where the K hidden units have non-overlapping receptive fields, and the only trainable parameters are the first layer weights $\{\boldsymbol{w}_l\}_{l=1}^K$, with $\boldsymbol{w}_l \in \mathbb{R}^{N/K}$ (a schematic view of this architecture is shown in Figure 1). The training dataset consists of $P = \alpha N$ input–label pairs $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)\}_{\mu=1}^P$, where we take the inputs $x_i^\mu \sim \mathcal{N}(0,1)$. For any choice of first-layer weights $\boldsymbol{w}$, we define the committee preactivation on pattern μ as

$$z^\mu(\boldsymbol{w}) \equiv \frac{1}{\sqrt{K}} \sum_{l=1}^K c_l \, \varphi\left(\sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} w_{li} \, x_{li}^\mu\right), \tag{1}$$

where we have used the indexing convention $x_{li}^\mu = x_{i+(l-1)N/K}^\mu$. The unknown teacher first-layer weights are drawn from a Gaussian distribution $w_{li}^\star \sim \mathcal{N}(0,1)$, the known second layer weights are chosen such that $\sum_{l=1}^K c_l = 0$ and $\sum_{l=1}^K c_l^2 = K$, and the output function depends on the task: for regression it is simply $f_\star(z) = z$, while for binary classification it is $f_\star(z) = \mathrm{sign}(z)$. The labels are generated by the teacher network, while the student has the same architecture and has to learn the first layer weights:

$$y^\mu = f_\star(z^\mu(\boldsymbol{w}^\star)), \qquad \hat{y}^\mu(\boldsymbol{w}) = z^\mu(\boldsymbol{w}). \tag{2}$$

Note how the student shares the same instance of the second layer weights as the teacher.

The student learns the weights $\boldsymbol{w}$ by minimizing the ℓ_2-regularized empirical loss

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{\alpha} \sum_{\mu=1}^P \ell(y^\mu, \hat{y}^\mu(\boldsymbol{w})) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{3}$$

The factor $1/\alpha$ ensures that the data-fitting and regularization terms are both of order N as α is varied. We define the set of regularized empirical risk minimizers (ERM) by

$$\widehat{\boldsymbol{w}}_{\mathrm{ERM}} \in \arg\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}). \tag{4}$$

To study its properties analytically, we introduce the Gibbs measure

$$p_\beta(\boldsymbol{w} \mid \mathcal{D}) = \frac{1}{Z_\beta} \exp[-\beta \mathcal{L}(\boldsymbol{w})], \qquad Z_\beta = \int d\boldsymbol{w} \, \exp[-\beta \mathcal{L}(\boldsymbol{w})]. \tag{5}$$

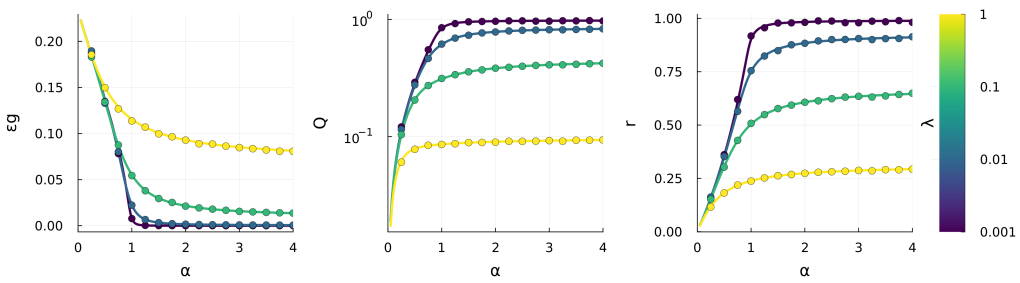

Figure 2: Generalization error and order parameters for regression. The RS theory predicts the test error ϵ_g, squared norm Q, and teacher–student overlap r as functions of sample complexity α and regularization strength λ. Curves are theoretical predictions; markers report finite-size ($N = 2000, K = 20$) GD simulations, averaged over 20 seeds. Error bars are not shown because they are very small.

In the $\beta \to \infty$ limit, this measure concentrates on the set of ERM estimators.

The Gibbs formulation of the ERM (5) can be used also to access its generalization ability. For a fresh Gaussian input $\boldsymbol{x}_\star$, let $y_\star$ be the teacher label and $\hat{y}_\star(\boldsymbol{w})$ the student's preactivation. In regression we use the test mean-squared error

$$\epsilon_g^{\text{regr}} = \mathbb{E}_{\boldsymbol{x}_\star}\mathbb{E}_{\mathcal{D}}\frac{1}{2}\left\langle \left(y_\star - \hat{y}_\star(\boldsymbol{w})\right)^2 \right\rangle_{\boldsymbol{w}|\mathcal{D}} . \tag{6}$$

Here $\langle\cdot\rangle_{\boldsymbol{w}|\mathcal{D}}$ denotes the average over the Gibbs measure (5). In classification we use the probability that teacher and student disagree

$$\epsilon_g^{\text{class}} = \mathbb{E}_{\boldsymbol{x}_\star}\mathbb{E}_{\mathcal{D}}\left\langle \Theta(-y_\star \hat{y}_\star(\boldsymbol{w}))\right\rangle_{\boldsymbol{w}|\mathcal{D}} . \tag{7}$$

The next section describes how the set of ERM estimators can be characterized by a finite set of macroscopic order parameters.

3 The ERM Estimator

3.1 Replica description and generalization error

We characterize typical ERM estimators in the proportional limit $N, P \to \infty$ with $\alpha = P/N$ fixed. The hidden-unit number is kept fixed while this thermodynamic limit is taken, and only afterwards we send $K \to \infty$. In this regime the quantity $\log Z_\beta / N$ concentrates, so its typical (or most probable) value is given by the quenched average, which is called the free entropy density

$$\phi_\beta = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\mathcal{D}} \log Z_\beta, \tag{8}$$

Here the notation $\mathbb{E}_{\mathcal{D}}$ denotes the average over the training set that depends on the particular realization of the inputs $\{\boldsymbol{x}^\mu\}_{\mu=1}^{\alpha N}$ and the teacher weights $\boldsymbol{w}^\star$.

We evaluate the free entropy using the replica method (Mézard et al., 1987), which expresses it in terms of a finite set of order parameters. These order parameters have a direct physical interpretation: they describe, after averaging over the dataset $\mathcal{D}$, the typical mutual overlap between independently sampled weights from $p_\beta(\cdot|\mathcal{D})$ and their typical alignment with the teacher. For hidden unit l, let us define:

$$Q_l^a = \mathbb{E}_{\mathcal{D}}\frac{K}{N}\sum_{i=1}^{N/K}(w_{li}^a)^2, \qquad q_l^{ab} = \mathbb{E}_{\mathcal{D}}\frac{K}{N}\sum_{i=1}^{N/K} w_{li}^a w_{li}^b , \qquad r_l^a = \mathbb{E}_{\mathcal{D}}\frac{K}{N}\sum_{i=1}^{N/K} w_{li}^a w_{li}^\star . \tag{9}$$

where $\boldsymbol{w}^a, \boldsymbol{w}^b \sim p_\beta(\cdot|\mathcal{D})$ are iid samples of the Gibbs distribution (5). Thus q_l^{ab} is the overlap between two Gibbs samples, Q_l^a is the squared norm of one sampled weight, and r_l is the teacher-student alignment. Since the hidden units are statistically equivalent and their receptive fields involve

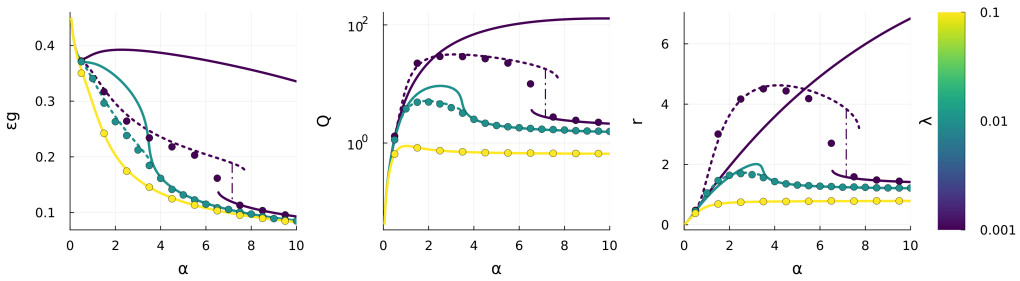

Figure 3: Classification order parameters. Solid curves denote RS solutions, while dashed curves denote 1RSB solutions. Markers report finite-size GD simulations for $N = 2000, K = 20$, averaged over 20 seeds. Error bars are not shown because they are very small.

disjoint, statistically independent subsets of the inputs, we can drop the hidden-unit index l: $q_l^{ab} = q^{ab}$ and $r_l = r$.

Although these quantities are averaged over $\mathcal{D}$, they are still random variables that depend on the draws from the Gibbs distribution, and could therefore have a non-trivial distribution. While it can be shown that both the norm and the teacher-student overlap have a trivial delta function distribution, this does not necessarily happen for the overlap q^{ab}. The remaining choice is how to approximate the distribution $P(q)$ of this quantity.

Under the replica symmetric (RS) ansatz this distribution is a single delta peak centered around a quantity q_0: $P(q) = \delta(q - q_0)$. In other words, all pairs of independently sampled students have the same typical overlap $q^{ab} = q_0$. The overlaps q^{ab} thus behave similarly to the norm Q and the teacher-student overlap r.

Replica symmetry breaking (RSB) ansatzes allow the same distribution to have more complicated shapes. The simplest one corresponds to the one-step replica symmetry breaking (1RSB) ansatz, which parametrizes the overlap distribution by two values $q_1 > q_0$: $P(q) = m\,\delta(q - q_0) + (1 - m)\,\delta(q - q_1)$. This can be interpreted by imagining samples that are organized into clusters: two samples in the same cluster have overlap q_1, while samples in different clusters have overlap q_0. The Parisi parameter m encodes the probability of observing the inter-cluster overlap q_0. More elaborate RSB ansatzes introduce additional overlap scales, and in the full-RSB limit the discrete set of peaks becomes a continuous probability distribution over overlaps. We discuss in the next subsection which of these ansatzes is selected, depending on the task and the particular region of the phase diagram examined. The value of the order parameters that parametrize the probability distributions is then obtained by solving a set of saddle point equations; see discussion in appendix B.

From these order parameters we can also access the generalization performance of the network. In particular, the definitions in Eqs. (6) and (7) can be evaluated from the joint Gaussian law of the teacher and student outputs on a fresh test input. In the $K \to \infty$ limit these errors become explicit functions of the order parameters Q and r:

$$\epsilon_g^{\text{class}} = \frac{1}{\pi} \arccos\left(\frac{D_0}{\sqrt{\Phi_\star\,(\Phi(Q) - \Phi(0))}}\right), \tag{10a}$$

$$\epsilon_g^{\text{regr}} = \frac{\Phi(Q) - \Phi(0) + \Phi_\star - 2D_0}{2}. \tag{10b}$$

The quantities $\Phi_\star$, D_0 and the function $\Phi(\bullet)$ appearing in (10) are written in terms of the infinite-width NNGP kernel

$$\mathcal{K}(Q_\star, Q, q) \equiv \int Dx\, Dy\, \varphi(\sqrt{Q_\star}x)\, \varphi\left(\frac{q}{\sqrt{Q_\star}}x + \sqrt{Q - \frac{q^2}{Q_\star}}y\right), \tag{11}$$

where Dx and Dy denote standard Gaussian measures. We use the shorthand $\Phi_\star = \mathcal{K}(1,1,1) - \mathcal{K}(1,1,0)$, $D_0 = \mathcal{K}(1,Q,r) - \mathcal{K}(1,Q,0)$, and $\Phi(q) = \mathcal{K}(Q,Q,q)$.

In the rest of the paper, we will focus on the square loss function

$$\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2. \tag{12}$$

Although for classification tasks other losses may achieve better performance, the square loss allows a Gaussian integral entering the replica calculation to be evaluated analytically (see appendix B for additional detail). It therefore gives a tractable high-dimensional landscape in which to study the properties of the ERM estimator. For a generic loss, the same integral must instead be evaluated numerically. We will also choose $\varphi(x) = \text{erf}(x)$ since the NNGP kernel in (11) can be evaluated analytically.

3.2 Regression: a replica symmetric phase diagram

First of all, we analyze the properties of the ERM estimator in the regression task, where the teacher output function is the identity. In this case, teacher and student architectures are fully matched, and we find that the full phase diagram is captured by a Replica Symmetric ansatz. Let us stress that a priori there is no theoretical guarantee that this should be the case: Replica Symmetry is only guaranteed for convex losses or in Bayesian-optimal settings, neither of which applies here.

The test error, norm and overlap with the teacher are shown in Figure 2 as a function of α for various values of λ. The behavior of the estimator is very similar to the Ridge Regression estimator: as long as the regularization is low enough, for $\alpha > 1$ the test error drops close to zero, and the teacher is almost fully recovered. For $\alpha < 1$ we find that the test error is non-monotonic in λ (not shown in the plot), while for $\alpha > 1$ it is monotonically decreasing with decreasing λ.
We then validate this RS ansatz by performing Gradient Descent experiments, and obtain an excellent match with our theory. More information on the numerics can be found in the appendix D.

3.3 Classification: Replica Symmetry Breaking

In the classification task, the ERM estimator displays a much richer phase diagram than its regression counterpart. The test error and the associated order parameters are shown in Fig. 3. At sufficiently low sample complexity, the system is well described by a replica symmetric (RS) saddle point for any value of the regularization strength λ.

As α increases, however, the shape of the landscape changes. For sufficiently small λ, an intermediate window of α opens in which the correct equilibrium description requires breaking the replica symmetry. Although the exact description of this phase involves full replica symmetry breaking (see Appendix B.3 for detail), we use a 1RSB approximation, which captures the relevant observables with only small corrections from higher RSB levels.

Finally, once α exceeds a threshold that depends explicitly on λ, the system exits the RSB phase and returns to a replica symmetric description, but of a qualitatively different character than the one found at small α. We refer to this final regime as the *recovery phase*. Here the test error drops sharply and the teacher-student cosine similarity $r/\sqrt{Q}$ increases markedly, indicating that the Gibbs measure has become dominated by a single, well-recovered minimum with substantially better generalization properties than the states populating the RSB region. The character of the transition connecting the RSB phase to the recovery phase itself depends sensitively on λ: for sufficiently small regularization (for instance $\lambda = 0.001$ in Fig. 3) the transition is discontinuous, with the order parameters jumping abruptly between the RSB branch and the recovery-phase branch, consistent with a first-order transition and an associated coexistence region. For larger values of λ the transition instead becomes continuous, with the RSB branch merging smoothly into the recovery phase as α is increased.

4 Hessian Spectrum in Minimizers

Having described the macroscopic properties of typical minimizers, we turn to computing their local curvature. Let $\mathcal{H}$ denote the Hessian of the data term in (3) evaluated at a typical trained configuration.

$$\mathcal{H} = \frac{1}{\alpha} \sum_{\mu=1}^{P} \frac{\partial^2 \ell(y^\mu, \hat{y}^\mu(\boldsymbol{w}))}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} \tag{13}$$

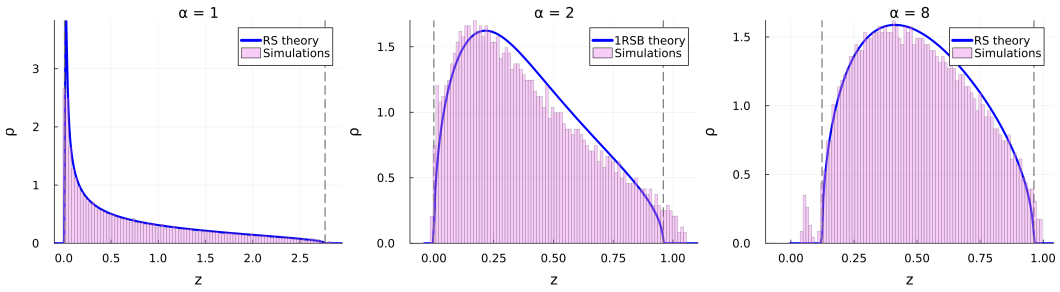

Figure 4: Empirical spectrum vs theoretical spectrum for $N = 2000$, $K = 20$ and $\lambda = 0.01$. The left panel shows regression; the middle and right panels show classification.

We are interested in computing the limiting spectral distribution of this matrix, $\rho_{\mathcal{H}}(z) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta(z - \lambda_i)$ where $\{\lambda_i\}_{i=1}^{N}$ are its eigenvalues. To this end we compute the Stieltjes transform $\mathcal{R}_{\mathcal{H}}(z) = \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\,(zI - \mathcal{H})^{-1}$ from which the limiting spectral distribution can be obtained via the Stieltjes inversion formula

$$\rho_{\mathcal{H}}(z) = \lim_{\eta\to 0} \frac{1}{\pi}\mathrm{Im}\,\mathcal{R}_{\mathcal{H}}(z - i\eta). \tag{14}$$

The Stieltjes transform can in turn be written as

$$\mathcal{R}_{\mathcal{H}}(z) = -\lim_{N\to\infty} \frac{2}{N}\frac{\partial}{\partial z} \ln \det(\mathcal{H} - zI)^{-1/2} \tag{15}$$

which, using a Gaussian integral representation of the determinant, allows us to use the replica method (Edwards & Jones, 1976), see Appendix C for additional detail.

In the case of the square loss function (12), the Stieltjes transform satisfies a cubic equation

$$1 + a_1(z)\mathcal{R}_{\mathcal{H}} + a_2(z)\mathcal{R}_{\mathcal{H}}^2 + a_3\mathcal{R}_{\mathcal{H}}^3 = 0, \tag{16}$$

where the coefficients $a_1(z), a_2(z), a_3$ are explicit functions of α, the activation kernel and its derivatives, and the order parameters and can be found in Appendix C.5. Consequently, the Hessian spectrum is determined by the same macroscopic state variables that were calculated above.

In Figure 4 we compare the theoretical spectrum with finite-N Gradient Descent simulations. The agreement is excellent in the RS phases, while a small discrepancy remains in the detailed spectral shape in the RSB phase. Since higher-order RSB corrections to the equilibrium order parameters are small (see Appendix B.3), we expect this mismatch to be more likely related to the extremely slow equilibration times characteristic of full-RSB landscapes (Sompolinsky & Zippelius, 1981; 1982; Cugliandolo & Kurchan, 1994).

Finally, note that in the high-α recovery phase, finite-N simulations show a set of K eigenvalues detaching from the left side of the continuous bulk (right panel of Fig. 4), with corresponding eigenvectors highly aligned to the teacher network. These modes are not captured by our limiting spectral density, since the large N limit is taken with fixed K and hence their spectral weight is $K/N \to 0$. We therefore exclude them when comparing the theoretical bulk edge with simulations, and report the edge of the continuous part of the spectrum. We also note that those isolated Hessian eigenvalues can have important dynamical consequences in other models. In phase retrieval, for example, an outlier eigenvalue can be associated with an eigenvector aligned with the planted signal, and its appearance or disappearance through a BBP transition is closely tied to whether gradient descent can escape marginal minima and recover the signal (Bonnaire et al., 2024).

5 Flatness is phase-dependent

We now ask which Hessian observables track generalization. We focus on three basic spectral summaries: the mean curvature

$$\bar{z} = \int x\,\rho_{\mathcal{H}}(x)\,\mathrm{d}x = \frac{1}{N}\mathrm{Tr}\,\mathcal{H}, \tag{17}$$

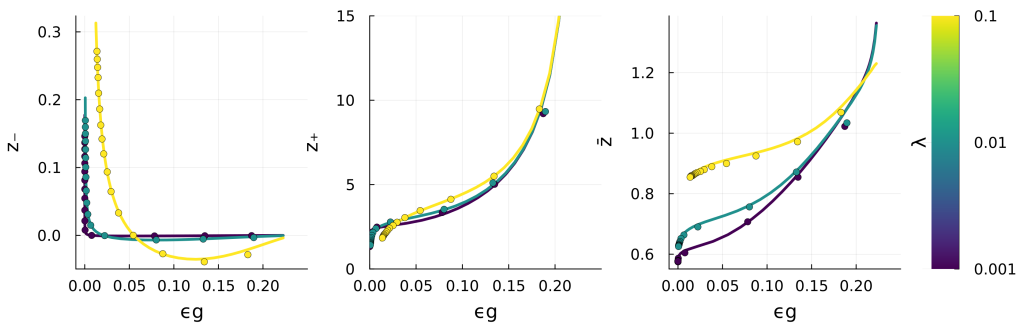

Figure 5: Flatness observables versus generalization error in regression. Markers are given by GD simulations with $N = 2000, K = 20$. In the phases where the empirical spectrum develops outliers, we show the smallest eigenvalue of the continuous bulk.

and the left and right spectral edges z_- and z_+. The right edge z_+ is a sharpness proxy, the trace measures the average curvature scale, and the left edge probes soft or marginal directions. Let us point out that our model possesses neither reparametrisation nor permutation symmetries, and therefore these flatness measures are not subject to the inconsistencies pointed out by Dinh et al. (2017). As we vary α, the generalization properties of the minimizers change. To check if a flatness-generalization relation holds in this setting, we plot these spectral observables as a function of the generalization error as α is varied, and visually inspect if a positive correlation holds.

Regression. We first address this question in the regression task. Figure 5 shows the relation between test error and the three flatness observables. Across the range of regularization strengths that we explored, the spectral mean and the right edge are positively correlated with the test loss. In other words, both the average curvature and the largest curvature provide a geometric quantity that correlates with generalization.
The left edge behaves differently. It correlates with the test error only in the overparametrized regime (where $\alpha < 1$) and with a degree that depends on the regularization λ, higher λ giving a higher correlation. In the underparametrized phase, instead, this relation breaks down, the left edge becoming anticorrelated with the test error.

Classification. A similar analysis can be performed for classification, using the 1RSB solution as the best available description of the relevant minimizers. In Fig. 6 we show the relation between spectral observables and test error for three values of the regularization λ, one for which the transition from the RSB phase to the recovery RS phase is continuous, and one for which it is discontinuous, and one for which the full phase diagram is RS. The resulting picture is qualitatively consistent with regression only within the first overparametrized RS phase, where both the spectral mean and the left and right edges remain reliably correlated with generalization. The correspondence breaks down in the other two phases: in the RSB regime, spectral observables are essentially uncorrelated with the generalization error, while in the recovery phase the correlation can in some cases reverse sign, with flatter minima associated with worse generalization. Hessian-based flatness measures are therefore phase-dependent diagnostics in classification, rather than universal proxies for predictive performance.

An interesting remark is that the left edge in the RSB phase is constant and equal to $-\lambda$. Indeed, the left edge cannot go below this value, since we are studying the Hessian of minimizers, and therefore the eigenvalues of the Hessian of the full loss (which contains a $+\lambda$ shift compared to the Hessian we are studying) cannot be negative. Our observation is thus that the RSB states predicted by our theory are marginal, a prediction which is matched by our simulations.

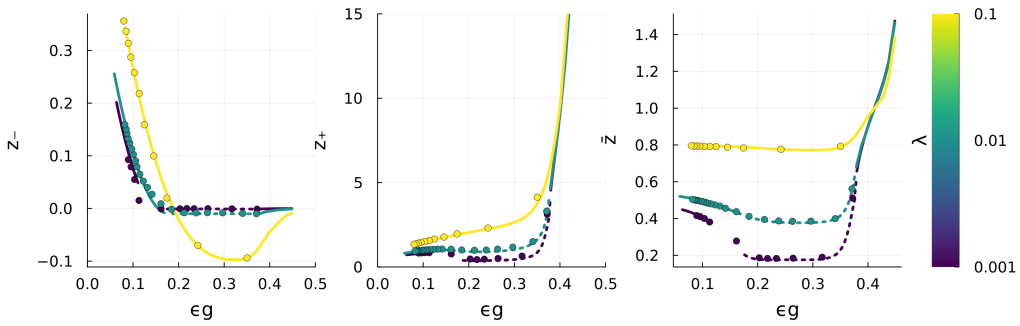

Figure 6: Flatness observables versus generalization error in classification. Solid curves are RS branches and dashed ones are 1RSB branches. Markers are given by GD simulations with $N = 2000, K = 20$. In the phases where the empirical spectrum develops outliers, we show the smallest eigenvalue of the continuous bulk.

6 Discussion and limitations

We have studied Hessian-based flatness measures of empirical-risk minimizers in a teacher-student tree committee machine model. The ERM estimators are characterized by a finite set of order parameters, which determine both their learning performance and the limiting Hessian spectrum around the learned solution. We focused on three standard flatness measures: the spectral mean, which captures the average curvature, and the left and right spectral edges, which probe the softest and sharpest directions. We validated the resulting predictions against finite-size gradient-descent simulations, finding quantitative agreement for all observables considered.

The central conclusion is that the flatness–generalization relation is phase-dependent rather than universal. In regression, where the phase diagram is replica symmetric, the spectral mean and right edge correlate with the test error across the regularizations we consider, while the left edge tracks it only for $\alpha < 1$. In classification the same observables predict generalization only in the overparametrized phase; in the intermediate RSB phase they are essentially uncorrelated with it, and in the high α replica symmetric recovery phase the correlation can reverse.

Two features of the RSB phase are worth isolating. First, the left edge is pinned at $-\lambda$ throughout this phase. Since the full regularized loss is minimized, its Hessian is positive semidefinite, so the data-term Hessian cannot have eigenvalues below $-\lambda$; our solution saturates this bound, indicating that these minimizers are marginal. This is reproduced by the simulations. Second, at large α the empirical Hessian develops a finite number of outlier eigenvalues to the left of the bulk. These are invisible to the limiting spectral density, which sees only $O(N)$ eigenvalues; their origin, and their relation to the class-structured Hessian outliers reported in deep classifiers, lies outside the present framework.

Finally, the context in which we analyze the flatness-generalization relation is slightly different from the one that is typically found in empirical works: there the relation is observed with a fixed dataset among different minimizers of the same loss landscape. In our setting, instead we modify the loss landscape by varying the dataset size to input dimension ratio. Strictly speaking, our results concern how curvature and generalization co-vary, and need not transfer directly to the fixed-data setting in which the relation is usually probed. Nevertheless, studying this relation as the landscape is varied provides a complementary perspective on the interplay between curvature and generalization. Furthermore, varying α has the advantage of sweeping across the RS, RSB, and recovery phases, which exposes the phase dependence we report. A similar analysis at fixed α, tracking distinct minimizers of a single landscape, would be needed to establish whether the same phase-conditioned picture holds within a landscape as well as across a family of them; we leave this to future work.

7 AI DISCLOSURE

In this work, we used generative AI tools for small steps in the replica computation, which were then independently verified. Additionally, we used generative AI tools for some of the plots, and for polishing the final version of this manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Martin Ahr, Michael Biehl, and Robert Urbanczik. Statistical physics and practical training of soft-committee machines. *The European Physical Journal B*, 10(3):583–588, 1999. doi: 10.1007/s100510050889.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. *arXiv preprint arXiv:2302.07011*, 2023.

Brandon L Annesi, Enrico Malatesta, and Francesco Zamponi. Exact full-rsb sat/unsat transition in infinitely wide two-layer neural networks. *SciPost Physics*, 18(4):118, 2025a. doi: 10.21468/SciPostPhys.18.4.118.

Brandon Livio Annesi, Dario Bocchi, and Chiara Cammarota. Overparametrization bends the landscape: Bbp transitions at initialization in simple neural networks. *arXiv preprint arXiv:2510.18435*, 2025b.

Kiana Asgari, Andrea Montanari, and Basil Saeed. Local minima of the empirical risk in high dimension: General theorems and convex examples. *arXiv preprint arXiv:2502.01953*, 2025.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. In *Advances in Neural Information Processing Systems 31*, pp. 3227–3238, 2018.

Carlo Baldassi, Enrico M Malatesta, and Riccardo Zecchina. Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations. *Physical review letters*, 123(17):170602, 2019.

Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1908636117.

Jean Barbier, Federica Gerace, Alessandro Ingrosso, Clarissa Lauditi, Enrico M. Malatesta, Gibbs Nwemadji, and Rodrigo Pérez Ortiz. Generalization performance of narrow one-hidden-layer networks in the teacher–student setting. *arXiv preprint arXiv:2507.00629*, 2025.

E Barkai, D Hansel, and I Kanter. Statistical mechanics of a multilayered neural network. *Physical review letters*, 65(18):2312, 1990.

E. Barkai, D. Hansel, and H. Sompolinsky. Broken symmetries in multilayered perceptrons. *Phys. Rev. A*, 45:4146–4161, Mar 1992. doi: 10.1103/PhysRevA.45.4146. URL https://link.aps.org/doi/10.1103/PhysRevA.45.4146.

Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643–656, 1995. doi: 10.1088/0305-4470/28/3/018.

Michael Biehl, Enno Schlösser, and Martin Ahr. Phase transitions in soft-committee machines. *Europhysics Letters*, 44(2):261–267, 1998. doi: 10.1209/epl/i1998-00466-6.

Tony Bonnaire, Giulio Biroli, and Chiara Cammarota. The role of the time-dependent hessian in high-dimensional optimization. *arXiv preprint arXiv:2403.02418*, 2024.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

L F Cugliandolo and J Kurchan. On the out-of-equilibrium relaxation of the sherrington-kirkpatrick model. *Journal of Physics A: Mathematical and General*, 27(17):5749, sep 1994. doi: 10.1088/0305-4470/27/17/011. URL https://doi.org/10.1088/0305-4470/27/17/011.

Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 2017.

S. F. Edwards and R. C. Jones. The eigenvalue spectrum of a large symmetric random matrix. *Journal of Physics A: Mathematical and General*, 9(10):1595–1603, 1976. doi: 10.1088/0305-4470/9/10/011.

Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 323–334. IEEE, 2022.

A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius. Storage capacity and learning algorithms for two-layer neural networks. *Phys. Rev. A*, 45:7590–7609, May 1992. doi: 10.1103/PhysRevA.45.7590. URL https://link.aps.org/doi/10.1103/PhysRevA.45.7590.

Andreas Engel and Christian Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

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.

Silvio Franz, Giorgio Parisi, Pierfrancesco Urbani, and Francesco Zamponi. Universal spectrum of normal modes in low-temperature glasses. *Proceedings of the National Academy of Sciences*, 112 (47):14539–14544, 2015.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241, 2019.

Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pp. 5–13, 1993. doi: 10.1145/168304.168306.

Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems*, volume 7, 1994.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1.

Haiping Huang and Yoshiyuki Kabashima. Origin of the computational hardness for learning with binary synapses. *Phys. Rev. E*, 90:052813, Nov 2014. doi: 10.1103/PhysRevE.90.052813.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL http://arxiv.org/abs/1609.04836.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

Zhenyu Liao and Michael W. Mahoney. Hessian eigenspectra of more realistic nonlinear models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987. doi: 10.1142/0271.

Andrea Montanari and Basil Saeed. Topological trivialization in non-convex empirical risk minimization. *arXiv preprint arXiv:2602.14969*, 2026.

Radford M. Neal. *Priors for Infinite Networks*, pp. 29–53. Springer New York, New York, NY, 1996. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra. *Journal of Machine Learning Research*, 21(252):1–64, 2020.

Giorgio Parisi. A sequence of approximated solutions to the s-k model for spin glasses. *Journal of Physics A: Mathematical and General*, 13(4):L115, 1980. doi: 10.1088/0305-4470/13/4/009. URL http://stacks.iop.org/0305-4470/13/i=4/a=009.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2798–2806, 2017.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. *arXiv preprint arXiv:2001.00939*, 2020.

Fabrizio Pittorino, Antonio Ferraro, Gabriele Perugini, Christoph Feinauer, Carlo Baldassi, and Riccardo Zecchina. Deep networks on toroids: removing symmetries reveals the structure of flat regions in the landscape geometry. *Journal of Statistical Mechanics: Theory and Experiment*, 2022 (11):114007, 2022. doi: 10.1088/1742-5468/ac9832.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52: 4225–4243, 1995. doi: 10.1103/PhysRevE.52.4225.

Mahalakshmi Sabanayagam, Freya Behrens, Urte Adomaityte, and Anna Dawid. Unveiling the hessian's connection to the decision boundary. *arXiv preprint arXiv:2306.07104*, 2023.

Levent Sagun, Utku Evci, V. Uğur Güney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Complex dynamics in simple neural networks: Understanding gradient flow in phase retrieval. In *Advances in Neural Information Processing Systems*, 2020.

Holm Schwarze and John A. Hertz. Statistical mechanics of learning in a large committee machine. In *Advances in Neural Information Processing Systems*, volume 5, pp. 523–530, 1992.

Neta Shoham, Liron Mor-Yosef, and Haim Avron. Flatness after all? *arXiv preprint arXiv:2506.17809*, 2025.

H. Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Phys. Rev. Lett.*, 47:359–362, Aug 1981. doi: 10.1103/PhysRevLett.47.359. URL https://link.aps.org/doi/10.1103/PhysRevLett.47.359.

H. Sompolinsky and Annette Zippelius. Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B*, 25:6860–6875, Jun 1982. doi: 10.1103/PhysRevB.25.6860. URL https://link.aps.org/doi/10.1103/PhysRevB.25.6860.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pp. 9636–9647. PMLR, 2020.

Harsh Vardhan, Hossein Taheri, and Arya Mazumdar. Flatness and generalization: Learning multi-index models with homogeneous neural networks. *arXiv preprint arXiv:2606.04429*, 2026.

Timothy L. H. Watkin, Albrecht Rau, and Michael Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499–556, 1993. doi: 10.1103/RevModPhys.65.499.

Christopher Williams. Computing with infinite networks. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/ae5e3ce40e0404a45ecacaaf05e5f735-Paper.pdf.

Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

Jacob A. Zavatone-Veth and Cengiz Pehlevan. Activation function dependence of the storage capacity of treelike neural networks. *Phys. Rev. E*, 103:L020301, Feb 2021. doi: 10.1103/PhysRevE.103.L020301. URL https://link.aps.org/doi/10.1103/PhysRevE.103.L020301.

Jacob A. Zavatone-Veth and Cengiz Pehlevan. On Neural Network Kernels and the Storage Capacity Problem. *Neural Computation*, 34(5):1136–1142, 04 2022. ISSN 0899-7667. doi: 10.1162/neco_a_01494. URL https://doi.org/10.1162/neco_a_01494.

Shuofeng Zhang, Isaac Reid, Guillermo Valle Pérez, and Ard Louis. Why flatness does and does not correlate with generalization for deep neural networks. *arXiv preprint arXiv:2103.06219*, 2021.

A Setting and objective

We will consider two identical architectures given by a one-hidden layer architecture with non-overlapping first layer weights (tree committee machines)

$$y^\mu = f_\star\left[\frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l\,\varphi\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} w^\star_{li}x^\mu_{li}\right)\right] \tag{18a}$$

$$\hat{y}^\mu(\boldsymbol{w}) = \frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l\,\varphi\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} w_{li}x^\mu_{li}\right) \equiv \frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l\,\varphi_l \tag{18b}$$

where $f_\star(x) = x$ for regression and $f_\star(x) = \mathrm{sign}(x)$ for classification. Moreover the $P = \alpha N$ input data $\boldsymbol{x}^\mu$ will be standard normal i.i.d. random variables as well as the weights of the teacher $w^\star_{li}$. The second layer weights c_l can be considered either to be fixed or sampled from a distribution. In the following we will assume that they satisfy the following conditions

$$\sum_l c_l = 0 \tag{19a}$$

$$\sum_l c_l^2 = K \tag{19b}$$

i.e. to have mean 0 and variance 1. We consider the empirical risk minimization

$$\hat{\boldsymbol{w}} = \text{argmin}_{\boldsymbol{w}} \left[\mathcal{L}(\boldsymbol{w})\right] \tag{20}$$

of a loss function $\mathcal{L}$ given by

$$\mathcal{L}(\boldsymbol{w}) = \sum_{\mu} \ell(y^\mu, \hat{y}^\mu(\boldsymbol{w})) + \frac{\lambda\alpha}{2}\|\boldsymbol{w}\|^2 \,. \tag{21}$$

y^μ and $\hat{y}^\mu$ are the output and the preactivation of the output of the teacher and student networks and λ represents the L2 regularization of the weights of the network. The ERM estimator $\hat{\boldsymbol{w}}$ can be obtained by studying the large β limit of the Gibbs measure (or posterior distribution)

$$p(\boldsymbol{w}|\mathcal{D}) = \frac{e^{-\beta\mathcal{L}(\boldsymbol{w})}}{Z_\beta} \tag{22}$$

where we have denoted by $\mathcal{D} = \{\boldsymbol{x}^\mu, y^\mu\}_{\mu=1}^{\alpha N}$ the dataset and by Z_β the partition function

$$Z_\beta = \int d\boldsymbol{w}\, e^{-\beta\mathcal{L}(\boldsymbol{w})} \tag{23}$$

Note that the loss defined here (21) differs from the one given in the main text by a factor α. This factor can be absorbed in the definition of the inverse temperature and does not give any contribution in the large β limit on which we focus.

B EQUILIBRIUM MEASURE

So far we did not specify how to compute the value of q, r and Q. This can be done by studying the equilibrium measure induced by the partition function defined in (23), in particular via the corresponding equilibrium free entropy

$$\phi^{\text{eq}} = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_{\mathcal{D}} \ln Z_\beta \,. \tag{24}$$

This quantity can be computed with standard methods, namely the replica method Mézard et al. (1987) which is based on the following identity

$$\ln Z_\beta = \lim_{n\to 0} \partial_n Z_\beta^n \tag{25}$$

We report here the outcome of the computation, which is reported elsewhere Baldassi et al. (2019). In the following we will denote by l the index that runs over the hidden units $l = 1, \dots, K$, and by a, b the indices running over the number of replicas $a, b = 1, \dots, n$. One finds that the equilibrium free entropy is obtained by a saddle point computation over a set of *order parameters* that can be conveniently condensed in K symmetric matrices of dimension $(n+1)\times(n+1)$ that we will call $Q_l^{\alpha\beta}$. The indices α, β will be indexed by $\alpha, \beta = 0, \dots, n$ for convenience. This matrix has a block structure of the type

$$\boldsymbol{Q}_l = \begin{pmatrix} 1 & r_l^a \\ r_l^a & q_l^{ab} \end{pmatrix} \tag{26}$$

where $q_l^{ab} \equiv \frac{K}{N}\sum_{i=1}^{N/K} w_{li}^a w_{li}^b$ with $a < b$ represents the overlap between the l-th hidden unit of two students extracted from the Gibbs measure (22), $q_l^{aa} = \frac{K}{N}\sum_{i=1}^{N/K}(w_{li}^a)^2$ is the typical squared norm of students weights and finally $r_l^a \equiv \frac{K}{N}\sum_{i=1}^{N/K} w_{li}^a w_{li}^\star$ is the overlap between the teacher and a student. Moreover $Q_l^{00} = 1$, which corresponds to the norm of the teacher, which is 1 in the large N limit by the central limit theorem.

One finds

$$\phi^{\text{eq}} = \lim_{n\to 0} \partial_n \,\text{extr}_{\boldsymbol{Q}} \left[G_{SI}^{\text{eq}}(\boldsymbol{Q}) + \alpha G_E^{\text{eq}}(\boldsymbol{Q})\right] \tag{27a}$$

$$G_{SI}^{\text{eq}}(\boldsymbol{Q}) \equiv \frac{1}{2K}\sum_l \ln\det\left(\boldsymbol{Q}_l\right) - \frac{\beta\alpha\lambda}{2K}\sum_l\sum_{a=1}^{n} Q_l^{aa} \tag{27b}$$

$$G_E^{\rm eq}(\boldsymbol{Q}) \equiv \ln \int \prod_l D_{\boldsymbol{Q}l} \boldsymbol{\lambda}_l \, e^{-\beta \sum_{a=1}^n \ell\left(f_\star\left[\frac{1}{\sqrt{K}} \sum_l c_l \, \varphi(\lambda_l^0)\right], \frac{1}{\sqrt{K}} \sum_l c_l \, \varphi(\lambda_l^a)\right)} \tag{27c}$$

where we have introduced the notation $D_{\boldsymbol{Q}}\boldsymbol{\lambda}$ to denote the integration over a $n+1$ dimensional Gaussian variable $\boldsymbol{\lambda}$ with vanishing mean vector and covariance matrix $\boldsymbol{Q}$. Note that for notational simplicity we have also rescaled the inverse temperature $\beta \to \alpha\beta$; this will not affect the computation of the free entropy of the system apart from an overall scale.

B.1 Large width limit

We now perform the limit of large number of hidden units $K \to \infty$ that we take only *after* the limits $P, N \to \infty$. Equivalently, we first determine the saddle points and evaluate the functions $G_{SI}^{\rm eq}$ and $G_E^{\rm eq}$ at those saddle point values, and only afterwards we consider their large K limit. Hence, although the number of hidden units is infinite, it remains of smaller order than both N and P. The limit can be done by noticing that each hidden unit accesses an uncorrelated portion of the input, so we expect the overlap matrix Q_l^{ab} to be independent of l

$$Q_l^{ab} = Q^{ab} \, . \tag{28}$$

We can exploit this property to substantially simplify both the entropic term $G_{SI}^{\rm eq}$ and energetic contributions $G_E^{\rm eq}$. The entropic term is straightforward and reads

$$G_{SI}^{\rm eq}(\boldsymbol{Q}) = \frac{1}{2} \ln \det(\boldsymbol{Q}) - \frac{\beta\alpha\lambda}{2} \sum_{a=1}^n Q^{aa} \, . \tag{29}$$

The energetic term is considerably more complicated to treat Baldassi et al. (2019); but it can be simplified by noticing that the arguments of the loss function $\frac{1}{\sqrt{K}} \sum_l c_l \, \varphi(\lambda_l^0)$ and $\frac{1}{\sqrt{K}} \sum_l c_l \, \varphi(\lambda_l^a)$ behave as Gaussian random variables. Following closely Baldassi et al. (2019); Annesi et al. (2025a) one has

$$G_E^{\rm eq}(\boldsymbol{Q}) = \ln \int D_{\boldsymbol{\mathcal{K}}}\boldsymbol{\lambda} \, e^{-\beta \sum_{a=1}^n \ell\left(f_\star[\lambda^0], \lambda^a\right)} \tag{30}$$

where the covariance matrix $\boldsymbol{\mathcal{K}}$ now reads

$$\boldsymbol{\mathcal{K}}_{ab} \equiv \int D_{\boldsymbol{Q}}\boldsymbol{\lambda} \, \varphi(\lambda_a)\varphi(\lambda_b) \, . \tag{31}$$

We have used the assumptions (19) on the second layer weights.

Note that (31) corresponds to the Neural Network Gaussian Process (NNGP) kernel that appears as the covariance matrix of the function implemented by a neural network at initialization in the infinite width limit and given two different inputs Neal (1996); Williams (1996); Zavatone-Veth & Pehlevan (2022).

B.2 RS ansatz

In order to perform the extremization and the limit $n \to 0$ in (27) we look for a saddle point matrix $\boldsymbol{Q}$ having a particular structure. The simplest ansatz is the *Replica Symmetric* one, which applied to the blocks of $\boldsymbol{Q}$ as in (26) reads

$$q^{ab} = Q \, \delta_{ab} + (1 - \delta_{ab}) \, q \tag{32a}$$

$$r^a = r \, . \tag{32b}$$

It is now an easy job to perform the $n \to 0$ limit. One obtains

$$\phi^{\rm eq} = \mathrm{extr}_{q,Q,r} \left[\mathcal{G}_{SI}^{\rm eq} + \alpha \mathcal{G}_E^{\rm eq}\right] \tag{33a}$$

$$\mathcal{G}_{SI}^{\rm eq} \equiv \frac{q - r^2}{2(Q-q)} + \frac{1}{2} \ln(Q-q) - \frac{\beta\alpha\lambda}{2} Q \tag{33b}$$

$$\mathcal{G}_E^{\rm eq} \equiv \int Dv Dx_1 \ln \int Dx_0 \, e^{-\beta \ell(y(v), \, \hat{y}(v, x_0, x_1))} \tag{33c}$$

where $y(v)$ and $\hat{y}(v, x_0, x_1)$ are defined as

$$y(v) \equiv f_\star\left(\sqrt{\Phi_\star} v\right) , \tag{34a}$$

$$\hat{y}(v, x_0, x_1) \equiv \frac{D_0}{\sqrt{\Phi_\star}} v + \sqrt{\Phi(Q) - \Phi(q)}\, x_0 + \sqrt{\Phi(q) - \Phi(0) - \frac{D_0^2}{\Phi_\star}}\, x_1 , \tag{34b}$$

The quantities $\Phi_\star$, D_0, $\Phi(\bullet)$ can be compactly expressed in terms of the NNGP kernel function

$$\mathcal{K}(Q_\star, Q, q) \equiv \int Dx Dy\, \varphi(\sqrt{Q_\star} x)\, \varphi\left(\frac{q}{\sqrt{Q_\star}} x + \sqrt{Q - \frac{q^2}{Q_\star}} y\right) \tag{35}$$

as

$$\Phi_\star = \mathcal{K}(1, 1, 1) - \mathcal{K}(1, 1, 0) \tag{36a}$$

$$D_0 = \mathcal{K}(1, Q, r) - \mathcal{K}(1, Q, 0) \tag{36b}$$

$$\Phi(q) = \mathcal{K}(Q, Q, q) \tag{36c}$$

B.3 k-STEPS REPLICA SYMMETRY BREAKING

The equilibrium free entropy in (27) can be obtained by searching for an extremizer in a more complex set of overlap matrices, which require the breaking of the replica symmetry Parisi (1980). We employ here a k-step Replica Symmetry Breaking (kRSB) ansatz for the overlap matrix q^{ab}

$$q^{ab} = q_0 + \sum_{s=0}^{k} (q_{s+1} - q_s) I_{ab}^{(n, m_s)} \tag{37}$$

where $I_{ab}^{(n,m)}$ is the (a, b) element of a block matrix of size $n \times n$ whose diagonal blocks have size $m \times m$; those $m \times m$ blocks are filled with ones whereas outside them all the elements are zero. In the previous equation we have also used the definitions

$$m_k = 1; \qquad q_{k+1} = Q\,. \tag{38}$$

For the teacher-student overlap r^a we keep the same ansatz as in (32). We remind here that the k-steps RSB matrix q^{ab} has $k + 2$ eigenvalues Annesi et al. (2025a) that read

$$\lambda_r \equiv \sum_{s=r}^{k} (q_{s+1} - q_s) m_s\,, \qquad d_r = n\left(\frac{1}{m_r} - \frac{1}{m_{r-1}}\right), \qquad r = -1\,, \dots\,, k \tag{39}$$

having defined $m_{-1} = n \to 0$ and $q_{-1} = 0$, $m_{-2} = \infty$. Using the Schur complement formula

$$\begin{aligned} \det \boldsymbol{Q} &= \left(1 - r^2\, \mathbb{I}^T \boldsymbol{q}^{-1} \mathbb{I}\right) \det \boldsymbol{q} = \left(1 - r^2 \sum_{ab} q_{ab}^{-1}\right) \det \boldsymbol{q} = \left(1 - r^2\, n \sum_{s=0}^{k} (q_{s+1}^{-1} - q_s^{-1}) m_s\right) \det \boldsymbol{q} \\ &= \left(1 - n \frac{r^2}{\lambda_0}\right) \det \boldsymbol{q} \end{aligned} \tag{40}$$

having used the fact that the set of k-step RSB matrices form a group. Using results exposed in Annesi et al. (2025a) the entropic term now reads

$$\mathcal{G}_{SI} = \frac{1}{2} \ln(Q - q_k) + \frac{q_0 - r^2}{2\lambda_0} + \sum_{s=0}^{k-1} \frac{1}{2 m_s} \ln\left(\frac{\lambda_s}{\lambda_{s+1}}\right) - \frac{\beta \alpha \lambda}{2} Q \tag{41}$$

The energetic term instead is

$$\mathcal{G}_E^{\text{eq}} \equiv \int Dv Dh\, f_0\left(v, \sqrt{\Phi(q_0) - \Phi(0) - \frac{D_0^2}{\Phi_\star}} h + \frac{D_0}{\sqrt{\Phi_\star}} v\right) \tag{42}$$

where f_0 is defined via the following recursion

$$f_k(v,h) = \ln \int Dz\, e^{-\beta \ell\left(f_\star(\sqrt{\Phi_\star}v),\, h+\sqrt{\Phi(Q)-\Phi(q_k)}\, z\right)} \tag{43a}$$

$$f_s(v,h) = \frac{1}{m_s} \ln \int Dz\, e^{m_s f_{s+1}\left(v,\, h+\sqrt{\Phi(q_{s+1})-\Phi(q_s)}\, z\right)}\,, \qquad s = k-1,\dots,0 \tag{43b}$$

Those expressions when evaluated to $k = 0, 1$ coincide with the RS (derived in section B.2) and 1RSB approximations.

Finally we point out that the equilibrium free entropy $\phi^{\rm eq}$ needs to be optimized with respect to the overlaps q_s, $s = 0,\dots,k$ and the breaking-points m_s, $s = 0,\dots,k-1$ as well as the squared norm Q and the teacher-student overlap r.

B.3.1 Large β limit

In the low temperature limit $\beta \to \infty$ the order parameter q_k and the breaking points m_s scale with β as

$$q_k = Q - \frac{\delta q}{\beta} \tag{44a}$$

$$m_s = \frac{\delta m_s}{\beta}\,, \qquad s = 0,\dots,k-1 \tag{44b}$$

This implies the following scaling in β for Φ_0

$$\Phi(Q) - \Phi(q_k) = \frac{\delta\Phi}{\beta} \tag{45a}$$

$\delta\Phi$ can be written as

$$\delta\Phi = \mathcal{K}^{(1,1)}(Q,Q,Q)\,\delta q \tag{46}$$

in terms of a generalized NNGP kernel function

$$\mathcal{K}^{(l,m)}(Q_\star,Q,q) \equiv \int DxDy\, \varphi^{(l)}(\sqrt{Q_\star}x)\varphi^{(m)}\left(\frac{q}{\sqrt{Q_\star}}x + \sqrt{Q-\frac{q^2}{Q_\star}}y\right) \tag{47}$$

where $\varphi^{(l)}$ denotes the l-th derivative of the activation function φ. Using this scaling the free energy $f_\beta = -\lim_{N\to\infty} \frac{1}{\beta N}\mathbb{E}_{\mathcal{D}} \ln Z_\beta$ is finite and can be written as

$$-f_\infty = \lim_{\beta\to\infty} \frac{\mathcal{G}_{SI}^{\rm eq}}{\beta} + \alpha \lim_{\beta\to\infty} \frac{\mathcal{G}_{E}^{\rm eq}}{\beta} \tag{48}$$

The entropic part reads in the limit as

$$\lim_{\beta\to\infty} \frac{\mathcal{G}_{SI}}{\beta} = \frac{q_0 - r^2}{2\delta\lambda_0} + \sum_{s=0}^{k-1} \frac{1}{2\delta m_s} \ln\left(\frac{\delta\lambda_s}{\delta\lambda_{s+1}}\right) - \frac{\alpha\lambda}{2}Q \tag{49}$$

where we have defined

$$\delta\lambda_s = \delta q + (Q - q_{k-1})\delta m_{k-1} + \sum_{r=s}^{k-2}(q_{r+1}-q_r)\delta m_r\,, \qquad s = 0\,,\dots,k-1 \tag{50a}$$

$$\delta\lambda_k = \delta q \tag{50b}$$

The energetic term instead reads

$$\lim_{\beta\to\infty} \frac{\mathcal{G}_{E}^{\rm eq}}{\beta} = \int DvDh\, f_0\left(v, \sqrt{\Phi(q_0) - \Phi(0) - \frac{D_0^2}{\Phi_\star}}h + \frac{D_0}{\sqrt{\Phi_\star}}v\right) \tag{51}$$

where now the f_0 satisfies the modified following recursion

$$f_k(v,h) = \max_z \left[-\frac{z^2}{2} - \ell(f_\star(\sqrt{\Phi_\star}v), h + \sqrt{\delta\Phi}\, z)\right] \tag{52a}$$

$$f_s(v,h) = \frac{1}{\delta m_s} \ln \int Dz\, e^{\delta m_s f_{s+1}\left(v,\, h+\sqrt{\Phi(q_{s+1})-\Phi(q_s)}\, z\right)}\,, \qquad s = k-1,\dots,0 \tag{52b}$$

B.3.2 SQUARE LOSS CASE

In the square loss case the previous recursion present in the energetic term can be fully solved analytically for generic β as all the integrals are Gaussian. Indeed one can guess the quantity

$$f_s(v,h) = A_s - \frac{B_s}{2}\left(h - f_\star\left(\sqrt{\Phi_\star}v\right)\right)^2, \qquad s = 0,\dots,k \tag{53}$$

is stable under recursion. The constants A_s and B_s can be derived accordingly. We start determining them by imposing the initial condition $s = k$ as in (43a); one finds

$$f_k(v,h) = -\frac{1}{2}\ln\left(1+\beta(\Phi(Q)-\Phi(q_k))\right) - \frac{\beta(h - f_\star\left(\sqrt{\Phi_\star}v\right))^2}{2(1+\beta(\Phi(Q)-\Phi(q_k)))} \tag{54a}$$

so that

$$A_k = -\frac{1}{2}\ln\left(1+\beta(\Phi(Q)-\Phi(q_k))\right) \tag{55a}$$

$$B_k = \frac{\beta}{1+\beta(\Phi(Q)-\Phi(q_k))} \tag{55b}$$

If we assume (53) for f_{s+1} one finds for f_s

$$\begin{aligned} f_s(v,h) &= \frac{1}{m_s}\ln\int Dz\, e^{m_s A_{s+1} - \frac{m_s B_{s+1}}{2}\left(h+\sqrt{\Phi(q_{s+1})-\Phi(q_s)}\,z - f_\star\left(\sqrt{\Phi_\star}v\right)\right)^2} \\ &= A_{s+1} - \frac{1}{2m_s}\ln\left(1+m_s B_{s+1}\left(\Phi(q_{s+1})-\Phi(q_s)\right)\right) - \frac{B_{s+1}\left(h - f_\star\left(\sqrt{\Phi_\star}v\right)\right)^2}{2\left(1+m_s B_{s+1}\left(\Phi(q_{s+1})-\Phi(q_s)\right)\right)} \end{aligned} \tag{56}$$

therefore A_s and B_s satisfy the recursion

$$A_s = A_{s+1} - \frac{1}{2m_s}\ln\left(1+m_s B_{s+1}\left(\Phi(q_{s+1})-\Phi(q_s)\right)\right) \tag{57a}$$

$$B_s = \frac{B_{s+1}}{1+m_s B_{s+1}\left(\Phi(q_{s+1})-\Phi(q_s)\right)} = \frac{1}{\frac{1}{B_{s+1}} + m_s\left(\Phi(q_{s+1})-\Phi(q_s)\right)} \tag{57b}$$

Those recursions can be solved. Let's start from B:

$$\begin{aligned} B_s &= \frac{1}{\frac{1}{B_{s+1}} + m_s\left(\Phi(q_{s+1})-\Phi(q_s)\right)} \\ &= \frac{1}{\frac{1}{B_{s+2}} + m_{s+1}\left(\Phi(q_{s+2})-\Phi(q_{s+1})\right) + m_s\left(\Phi(q_{s+1})-\Phi(q_s)\right)} \\ &\vdots \\ &= \frac{\beta}{1+\beta\chi_s}, \qquad s = 0,\dots,k \end{aligned} \tag{58}$$

having defined

$$\chi_s = \sum_{r=s}^{k}\left(\Phi(q_{r+1})-\Phi(q_r)\right)m_r, \qquad s = 0,\dots,k \tag{59}$$

Correspondingly the coefficients A_s are given by

$$\begin{aligned} A_s &= A_{s+1} - \frac{1}{2m_s}\ln\left(1+\frac{\beta m_s}{1+\beta\chi_{s+1}}\left(\Phi(q_{s+1})-\Phi(q_s)\right)\right) = A_{s+1} - \frac{1}{2m_s}\ln\left(\frac{1+\beta\chi_s}{1+\beta\chi_{s+1}}\right) \\ &= -\sum_{r=s}^{k}\frac{1}{2m_r}\ln\left(\frac{1+\beta\chi_r}{1+\beta\chi_{r+1}}\right) = -\sum_{r=s}^{k-1}\frac{1}{2m_r}\ln\left(\frac{1+\beta\chi_r}{1+\beta\chi_{r+1}}\right) - \frac{1}{2}\ln(1+\beta\chi_k), \qquad s = 0,\dots,k \end{aligned} \tag{60}$$

Therefore the energetic term reads

$$\begin{aligned}\mathcal{G}_E^{\rm eq} &\equiv \int DvDh\, f_0\left(v, \sqrt{\Phi(q_0)-\Phi(0)-\frac{D_0^2}{\Phi_\star}}h+\frac{D_0}{\sqrt{\Phi_\star}}v\right)\\ &= A_0-\frac{B_0}{2}\int DvDh\left(\sqrt{\Phi(q_0)-\Phi(0)-\frac{D_0^2}{\Phi_\star}}h+\frac{D_0}{\sqrt{\Phi_\star}}v-f_\star\left(\sqrt{\Phi_\star}v\right)\right)^2\end{aligned} \tag{61}$$

The remaining integrals over v and h can be solved as well, but the result depends on the task i.e. on $f_\star$. For the classification case one finds

$$\begin{aligned}\mathcal{G}_E^{\rm eq} &= A_0-\frac{B_0}{2}\int DvDh\left(\sqrt{\Phi(q_0)-\Phi(0)-\frac{D_0^2}{\Phi_\star}}h+\frac{D_0}{\sqrt{\Phi_\star}}v-\mathrm{sign}(v)\right)^2\\ &= A_0-\frac{B_0}{2}\left[\Phi(q_0)-\Phi(0)-\frac{D_0^2}{\Phi_\star}+\frac{D_0^2}{\Phi_\star}+1-2\frac{D_0}{\sqrt{\Phi_\star}}\int Dv\, v\,\mathrm{sign}(v)\right]\\ &= A_0-\frac{B_0}{2}\left[1+\Phi(q_0)-\Phi(0)-2\sqrt{\frac{2}{\pi}}\frac{D_0}{\sqrt{\Phi_\star}}\right]\end{aligned} \tag{62}$$

or

$$\mathcal{G}_E^{\rm eq} = -\frac{\beta}{2(1+\beta\chi_0)}\left(1+\Phi(q_0)-\Phi(0)-2D_0\sqrt{\frac{2}{\pi\Phi_\star}}\right)-\frac{1}{2}\ln(1+\beta\chi_k)-\sum_{s=0}^{k-1}\frac{1}{2m_s}\ln\left(\frac{1+\beta\chi_s}{1+\beta\chi_{s+1}}\right), \tag{63}$$

In the regression instead

$$\begin{aligned}\mathcal{G}_E^{\rm eq} &= A_0-\frac{B_0}{2}\int DvDh\left(\sqrt{\Phi(q_0)-\Phi(0)-\frac{D_0^2}{\Phi_\star}}h+\left(\frac{D_0}{\sqrt{\Phi_\star}}-\sqrt{\Phi_\star}\right)v\right)^2\\ &= A_0-\frac{B_0}{2}\left[\Phi(q_0)-\Phi(0)+\Phi_\star-2D_0\right]\\ &= -\frac{\beta}{2(1+\beta\chi_0)}\left(\Phi(q_0)-\Phi(0)+\Phi_\star-2D_0\right)-\frac{1}{2}\ln(1+\beta\chi_k)-\sum_{s=0}^{k-1}\frac{1}{2m_s}\ln\left(\frac{1+\beta\chi_s}{1+\beta\chi_{s+1}}\right)\end{aligned} \tag{64}$$

Because the free entropy is analytic in terms of q_s, $s=0,\dots,k$, m_s, $s=0,\dots,k-1$ and Q, r it is easy to optimize it for generic k. In Figure 7 we show the result of this optimization in the classification case, by comparing $k=0$ to the cases $k=1$ and $k=20$. In particular, we show that the squared norm Q changes substantially when going from the $k=0$ RS solution to the $k=1$ RSB solution at large β, while further RSB steps produce only comparatively small corrections to the 1RSB value.

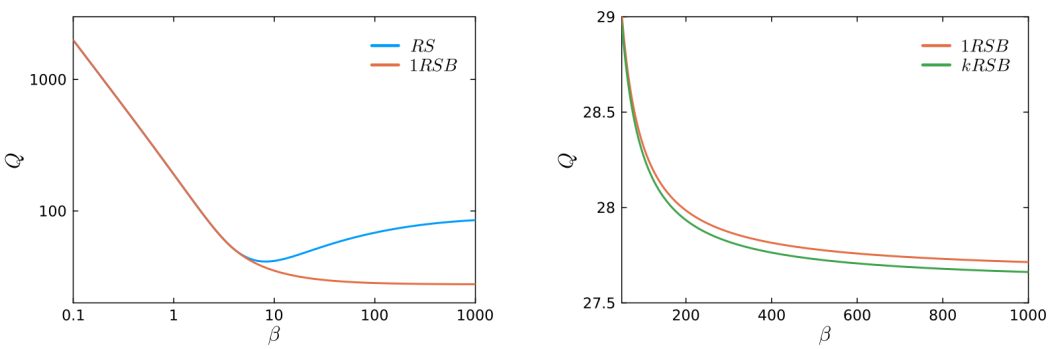

Figure 7: Classification case with square loss function. Dependence of the squared norm Q on the inverse temperature β for $\lambda = 0.001$ and $\alpha = 5$. Left panel: comparison between the replica symmetric solution and the 1RSB solution. Right panel: comparison between the 1RSB solution and the finite-step k-RSB approximation with $k = 20$. In the k-RSB solution we have optimized also on the breaking points m_s, $s = 0, \dots, k-1$. In both cases the curves coincide at high temperature and separate as β increases.

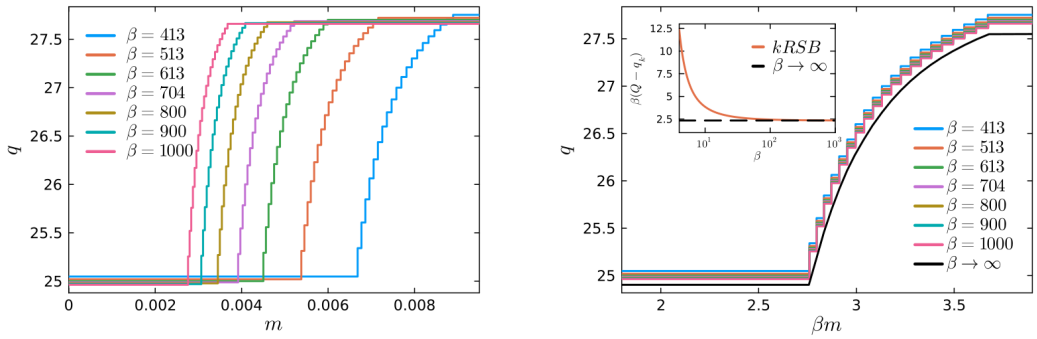

Figure 8: Classification case with square loss function. Overlap function q_s vs breaking points m_s for $\alpha = 5$, $\lambda = 0.001$ and several values of β (left panel). On the right panel we have plotted the same data of the left panel vs $\beta m_s \equiv \delta m_s$. As predicted in (44) this is the right scaling of the breaking-points for large β. In black we also depict the exact solution in the limit $\beta \to \infty$. In the inset we show a numerical verification of (44), plotting $\beta(Q - q_k)$ vs β and showing that it approaches a finite limit δq (black dashed line).

In the left panel of Figure 8 we instead show the overlap q vs the breaking point value m. In the right panel we show a validation of the large beta scalings introduced in (44).

The analytical large β limit expression of the free entropy can be obtained by applying the scalings in (44) or using the same approach above to solve the recursion in (52). One finds for classification the following expression

$$\lim_{\beta\to\infty} \frac{\mathcal{G}_E^{\rm eq}}{\beta} = -\frac{1}{2(1+\delta\chi_0)}\left(1+\Phi(q_0)-\Phi(0)-2D_0\sqrt{\frac{2}{\pi\Phi_\star}}\right) - \sum_{s=0}^{k-1}\frac{1}{2\delta m_s}\ln\left(\frac{1+\delta\chi_s}{1+\delta\chi_{s+1}}\right), \tag{65}$$

while for regression

$$\lim_{\beta\to\infty} \frac{\mathcal{G}_E^{\rm eq}}{\beta} = -\frac{1}{2(1+\delta\chi_0)}\left(\Phi(q_0)-\Phi(0)+\Phi_\star-2D_0\right) - \sum_{s=0}^{k-1}\frac{1}{2\delta m_s}\ln\left(\frac{1+\delta\chi_s}{1+\delta\chi_{s+1}}\right). \tag{66}$$

In the previous expressions we have defined

$$\delta\chi_s = \delta\Phi + (\Phi(Q) - \Phi(q_{k-1}))\,\delta m_{k-1} + \sum_{r=s}^{k-2} (\Phi(q_{r+1}) - \Phi(q_r))\,\delta m_r\,, \qquad s = 0\,, \dots\,, k-1 \tag{67}$$

$$\delta\chi_k = \delta\Phi\,. \tag{68}$$

In Figure 9 we show the squared norm Q as a function of α in the classification case comparing $k = 1$ and $k = 20$ steps of RSB in the large β limit. We also present some representative plots for the overlap q in the right panel.

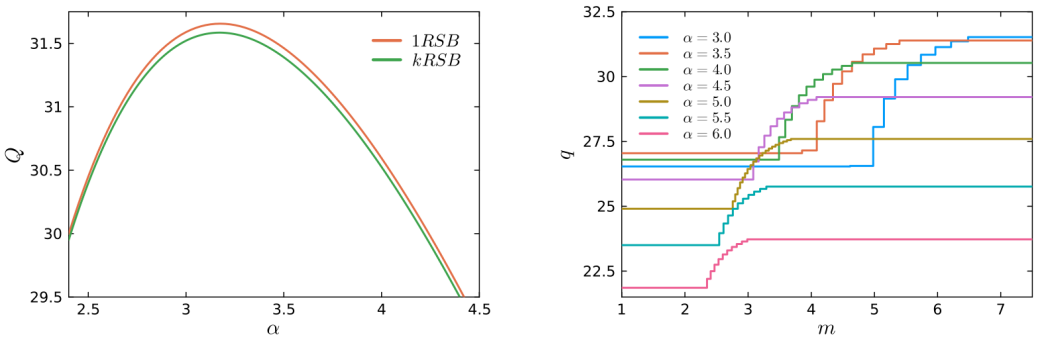

Figure 9: Classification case with square loss function. Left panel: squared norm Q versus α in the large β regime, for $\lambda = 0.001$. We compare the 1RSB (orange curve) and k-RSB with $k = 20$ (green). In the k-RSB solution we have optimized also on the breaking points δm_s, $s = 0, \dots, k-1$; we therefore expect the green curve to be a good proxy of the full-RSB line (i.e. $k \to \infty$). More than one step of RSB are therefore only slightly influencing the squared norm. We point out here that other order parameters are for all purposes left unchanged. Right panel: overlap vs breaking-points for several values of α. The maximum value achieved by the overlap is Q.

B.4 Generalization and training error

In classification tasks ($f_\star = \text{sign}$) the standard measure of predictive performance after training is the generalization error. It is defined by computing the probability of wrongly predicting a label $y_\star$ which has been generated by the teacher network by sampling a new Gaussian input $\boldsymbol{x}_\star$ (i.e. not in the training set), i.e.

$$\epsilon_g = \mathbb{E}_{\boldsymbol{x}_\star} \mathbb{E}_{\mathcal{D}} \left\langle \Theta\left(-y_\star \hat{y}_\star(\boldsymbol{w})\right)\right\rangle_{\boldsymbol{w}|\mathcal{D}} \tag{69}$$

where we remind that $\langle \bullet \rangle_{\boldsymbol{w}|\mathcal{D}}$ denotes the average over the Gibbs measure (22) and $\mathbb{E}_{\mathcal{D}}$ is the average over the realization of the dataset. $\hat{y}_\star(\boldsymbol{w})$ denotes the preactivation of the student's output corresponding to the test input $\boldsymbol{x}_\star$. This quantity can be easily computed again by using the replica method. In the $K \to \infty$ limit, in particular, one finds

$$\epsilon_g = \frac{1}{\pi} \arccos\left(\frac{D_0}{\sqrt{\Phi_\star(\Phi(Q) - \Phi(0))}}\right) \tag{70}$$

In regression we instead measure the test loss. In the $K \to \infty$ limit one has

$$\epsilon_g = \mathbb{E}_{\boldsymbol{x}_\star} \mathbb{E}_{\mathcal{D}} \frac{1}{2} \left\langle (y_\star - \hat{y}_\star(\boldsymbol{w}))^2 \right\rangle_{\boldsymbol{w}|\mathcal{D}} = \frac{\Phi(Q) - \Phi(0) + \Phi_\star - 2D_0}{2} \tag{71}$$

The training loss and the training error respectively relevant in regression and classification tasks can be found similarly. For example the training error is given by the following expression

$$\epsilon_t = \mathbb{E}_{\mathcal{D}} \left\langle \frac{1}{\alpha N} \sum_{\mu} \Theta\left(-y\hat{y}(\boldsymbol{w})\right) \right\rangle_{\boldsymbol{w}|\mathcal{D}} = \left\langle \Theta\left(-y(v)\,\hat{y}(v, x_0, x_1, x_{01})\right)\right\rangle_{\beta}\,. \tag{72}$$

C REPLICA COMPUTATION OF THE SPECTRUM OF THE HESSIAN

C.1 GIBBS FORMULATION OF THE SPECTRUM OF THE HESSIAN

Consider $\mathcal{H}$ to be a generic random matrix of size $N \times N$. Its limiting spectral distribution is defined as

$$\rho_{\mathcal{H}}(z) = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \delta(z-\lambda_i) \tag{73}$$

where we have denoted by λ_i, $i \in [N]$ the eigenvalues of $\mathcal{H}$. A fundamental quantity related to the empirical spectral distribution $\rho_{\mathcal{H}}(z)$ is the so-called Stieltjes transform, which is the normalized trace of the resolvent of $\mathcal{H}$

$$\mathcal{R}_{\mathcal{H}}(z) \equiv \lim_{N\to\infty} \frac{1}{N}\mathrm{Tr}\,(zI-\mathcal{H})^{-1} = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{z-\lambda_i} = \int d\lambda\, \frac{\rho_{\mathcal{H}}(\lambda)}{z-\lambda} \tag{74}$$

The previous expression can be inverted as

$$\rho_{\mathcal{H}}(z) = \lim_{\varepsilon\to 0^+}\frac{1}{\pi}\mathrm{Im}\mathcal{R}_{\mathcal{H}}(z-i\varepsilon) \tag{75}$$

The Stieltjes transform can be computed by using the replica method, since it can be written as the derivative of a quantity that plays the role of a "free entropy" $\phi_{\mathcal{H}}(z)$

$$\mathcal{R}_{\mathcal{H}}(z) = \frac{\partial}{\partial z}\phi_{\mathcal{H}}(z) \tag{76}$$

which is defined as

$$\phi_{\mathcal{H}}(z) = -\frac{2}{N}\ln \mathcal{Z}_{\mathcal{H}}(z) \tag{77a}$$

$$\mathcal{Z}_{\mathcal{H}}(z) = \det(\mathcal{H}-zI)^{-1/2} = \int\prod_{i=1}^{N}\frac{du_i}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\sum_{ij}u_i(\mathcal{H}_{ij}-z\delta_{ij})u_j} \tag{77b}$$

$\mathcal{Z}_{\mathcal{H}}(z)$ plays the role of a partition function. Notice that similarly the cumulative distribution of $\mathcal{H}$ can be obtained by computing the imaginary part of $\phi_{\mathcal{H}}$

$$\mathcal{C}_{\mathcal{H}}(x) = \int_{-\infty}^{x} dz\, \rho_{\mathcal{H}}(z) = \frac{1}{\pi}\mathrm{Im}\phi_{\mathcal{H}}(x)\,. \tag{78}$$

The previous identities are valid for a generic $N \times N$ random matrix $\mathcal{H}$. In the following we will be interested in computing the spectrum of the Hessian of the loss function of the teacher-student tree committee machine model as defined in section A, for a typical weight sampled from the Gibbs measure (22). This means that we need to compute

$$-2\phi(z) \equiv \lim_{N\to\infty}\mathbb{E}_{\mathcal{D}}\langle\phi_{\mathcal{H}}(z)\rangle_{\boldsymbol{w}|\mathcal{D}} = -\lim_{N\to\infty}\frac{2}{N}\mathbb{E}_{\mathcal{D}}\langle\ln\mathcal{Z}_{\mathcal{H}}(z)\rangle_{\boldsymbol{w}|\mathcal{D}} \tag{79}$$

where $\langle\bullet\rangle_{\boldsymbol{w}|\mathcal{D}}$ denotes the average over the Gibbs measure (22) and $\mathbb{E}_{\mathcal{D}}$ is the average over the realization of the dataset.

The Hessian

$$\alpha\mathcal{H} = \sum_{\mu=1}^{P}\frac{\partial^2\ell(y^\mu,\hat{y}^\mu(\boldsymbol{w}))}{\partial\boldsymbol{w}\partial\boldsymbol{w}^T} \tag{80}$$

is in our case an $N \times N$ matrix having the following block structure

$$\alpha\mathcal{H}_{li,l'i'} = \frac{1}{N}\sum_{\mu}\ell''(y^\mu,\hat{y}^\mu)c_l c_{l'}\varphi_l'\varphi_{l'}' x_{li}^\mu x_{l'i'}^\mu + \delta_{ll'}\frac{\sqrt{K}}{N}\sum_{\mu}\ell'(y^\mu,\hat{y}^\mu)c_l\varphi_l'' x_{li}^\mu x_{li'}^\mu\,. \tag{81}$$

where we have used the notation $\varphi_l = \varphi\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} w_{li}x_{li}^\mu\right)$ and we have denoted by a prime the derivative of the loss function with respect to its second argument $\hat{y}$. Notice that we have not included in this definition the second derivative with respect to $\boldsymbol{w}$ of the L2 regularization appearing in the loss $\mathcal{L}(\boldsymbol{w})$ in (3). Since the regularization term is quadratic in $\boldsymbol{w}$ this provides only a shift to the spectrum of the Hessian depending on the value of λ. Notice that the Hessian that we compute here differs from the one in the main text by a scaling factor α; this factor can be taken into account a posteriori by appropriate rescaling of the spectrum: $\rho_{\mathcal{H}}(z) = \alpha\rho_{\alpha\mathcal{H}}(\alpha z)$.

C.2 REPLICA TRICKS

In order to average over the input data, we use two replica tricks, one for $1/Z_\beta = \lim_{n\to 0} Z_\beta^{n-1}$ and $\ln \mathcal{Z}_{\alpha\mathcal{H}} = \lim_{s\to 0} \partial_s \mathcal{Z}^s_{\alpha\mathcal{H}}$. Notice indeed the similarity of our computation with the standard Franz-Parisi potential Franz & Parisi (1995). By analogy the variables u_{li} are acting as "fake" weights whereas the variable z of the spectral density $\rho_{\alpha\mathcal{H}}(z)$ acts like a regularization over those fake weights.

$$\phi(z) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_{\mathcal{D}} \langle \ln \mathcal{Z}_{\alpha\mathcal{H}}(z) \rangle_{\boldsymbol{w}|\mathcal{D}} = \lim_{N\to\infty} \lim_{\substack{n\to 0 \\ s\to 0}} \frac{\partial_s \tilde{\phi}}{N} \tag{82}$$

where

$$\tilde{\phi} \equiv \mathbb{E}_{\mathcal{D}} \int \prod_{a=1}^{n} d\boldsymbol{w}^a \, e^{-\beta \sum_{a=1}^n \mathcal{L}(\boldsymbol{w}^a)} \, \mathcal{Z}^s_{\mathcal{H}}(z) \tag{83}$$

In the following we will denote by $a, b \in [n]$ whereas $c, d \in [s]$. Using the expression of the Hessian in (81), the quadratic term in the partition function in (77b) is given by

$$\begin{aligned} \alpha \sum_{li,l'i'} u_{li} \mathcal{H}_{li,l'i'} u_{l'i'} = \sum_\mu \ell''(y^\mu, \hat{y}^\mu) \left(\frac{1}{\sqrt{K}} \sum_l c_l \varphi_l' \sqrt{\frac{K}{N}} \sum_i u_{li} x^\mu_{li} \right)^2 \\ + \sum_\mu \ell'(y^\mu, \hat{y}^\mu) \frac{1}{\sqrt{K}} \sum_l c_l \left(\sqrt{\frac{K}{N}} \sum_i u_{li} x^\mu_{li} \right)^2 \varphi_l'' \end{aligned} \tag{84}$$

We therefore found that the dependence on the disorder in the right hand side of (83) appears via the three variables

$$\eta^\mu_l = \sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} w^\star_{li} x^\mu_{li} \tag{85a}$$

$$\lambda^\mu_{la} = \sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} w^a_{li} x^\mu_{li} \tag{85b}$$

$$h^\mu_{lc} = \sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} u^c_{li} x^\mu_{li} \tag{85c}$$

which are jointly Gaussian random variables. Equivalently we can enforce those definitions via delta functions insertions and represent those delta functions via their Fourier representation. The average over the input data reads

$$\begin{aligned} &\prod_{li} \prod_{\mu} \mathbb{E}_{x^\mu_{li}} e^{-i\sqrt{\frac{K}{N}} x^\mu_{li} \left(\hat{\eta}^\mu_l w^\star_{li} + \sum_a \hat{\lambda}^\mu_{la} w^a_{li} + \sum_c \hat{h}^\mu_{lc} u^c_{li} \right)} = \prod_{li} \prod_{\mu} e^{-\frac{K}{2N} \left(\hat{\eta}^\mu_l w^\star_{li} + \sum_a \hat{\lambda}^\mu_{la} w^a_{li} + \sum_c \hat{h}^\mu_{lc} u^c_{li} \right)^2} \\ &= \prod_{\mu l} e^{-\frac{1}{2}(\hat{\eta}^\mu_l)^2 - \frac{1}{2}\sum_{ab} q^{ab}_l \hat{\lambda}^\mu_{la} \hat{\lambda}^\mu_{lb} - \frac{1}{2}\sum_{cd} p^{cd}_l \hat{h}^\mu_{lc} \hat{h}^\mu_{ld} - \sum_{ac} t^{ac}_l \hat{\lambda}^\mu_{la} \hat{h}^\mu_{lc} - \hat{\eta}^\mu_l \sum_a r^a_l \hat{\lambda}^\mu_{la} - \hat{\eta}^\mu_l \sum_c o^c_l \hat{h}^\mu_{lc}} \end{aligned} \tag{86}$$

where we have introduced the following overlap quantities

$$q^{ab}_l \equiv \frac{K}{N} \sum_{i=1}^{N/K} w^a_{li} w^b_{li} \,, \qquad r^a_l \equiv \frac{K}{N} \sum_{i=1}^{N/K} w^\star_{li} w^a_{li} \tag{87a}$$

$$p^{cd}_l \equiv \frac{K}{N} \sum_{i=1}^{N/K} u^c_{li} u^d_{li} \,, \qquad t^{ac}_l \equiv \frac{K}{N} \sum_{i=1}^{N/K} w^a_{li} u^c_{li} \,, \qquad o^c_l \equiv \frac{K}{N} \sum_{i=1}^{N/K} w^\star_{li} u^c_{li} \tag{87b}$$

and we have also used the fact that $\frac{K}{N}\sum_i (w^\star_{li})^2 = 1$ for large N thanks to the central limit theorem. The previous overlaps can be enforced via delta functions. In the end one finds the free entropy to be

$$\tilde{\phi} = \int \prod_{abl} \frac{dq^{ab}_l d\hat{q}^{ab}_l}{2\pi} \prod_{acl} \frac{dt^{ac}_l d\hat{t}^{ac}_l}{2\pi} \prod_{cdl} \frac{dp^{cd}_l d\hat{p}^{cd}_l}{2\pi} \prod_{al} \frac{dr^a_l d\hat{r}^a_l}{2\pi} \prod_{cl} \frac{do^c_l d\hat{o}^c_l}{2\pi} e^{N(G_I + G_S + \alpha G_E)} \tag{88}$$

where the interaction, entropic and energetic terms are the following:

$$G_I = -\frac{1}{2K}\sum_{abl} q_l^{ab}\hat{q}_l^{ab} - \frac{1}{K}\sum_{acl} t_l^{ac}\hat{t}_l^{ac} - \frac{1}{2K}\sum_{cdl}\hat{p}_l^{cd}p_l^{cd} - \frac{1}{K}\sum_{al} r_l^a\hat{r}_l^a - \frac{1}{K}\sum_{cl} o_l^c\hat{o}_l^c + \frac{z}{2K}\sum_{lc} p_l^{cc} \tag{89a}$$

$$G_S = \frac{1}{K}\log \mathbb{E}_{\boldsymbol{w}^\star}\int\prod_{la} dw_l^a \prod_{lic}\frac{du_{li}^c}{\sqrt{2\pi}} e^{\frac{1}{2}\sum_{abl}\hat{q}_l^{ab}w_l^a w_l^b + \frac{1}{2}\sum_{cdl}\hat{p}_l^{cd}u_l^c u_l^d + \sum_{acl}\hat{t}_l^{ac}w_l^a u_l^c + \sum_{al}\hat{r}_l^a w_l^\star w_l^a + \sum_{cl}\hat{o}_l^c u_l^c w_l^\star} \tag{89b}$$

$$\begin{aligned} G_E = \log\int\prod_l \frac{d\eta_l d\hat{\eta}_l}{2\pi}\prod_{al}\frac{d\lambda_{la}d\hat{\lambda}_{la}}{2\pi}\prod_{cl}\frac{d\hat{h}_{lc}dh_{lc}}{2\pi} e^{i\sum_l \eta_l\hat{\eta}_l + i\sum_{la}\lambda_{la}\hat{\lambda}_{la} + i\sum_{lc}h_{lc}\hat{h}_{lc}} \\ \times e^{-\beta\sum_a \ell(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_a))} \\ \times e^{-\frac{1}{2}\sum_l\left(\hat{\eta}_l^2 + \sum_{ab} q_l^{ab}\hat{\lambda}_{la}\hat{\lambda}_{lb} + \sum_{cd} p_l^{cd}\hat{h}_{lc}\hat{h}_{ld} + 2\sum_{ac} t_l^{ac}\hat{\lambda}_{la}\hat{h}_{lc} + 2\hat{\eta}_l\sum_a r_l^a\hat{\lambda}_{la} + 2\hat{\eta}_l\sum_c o_l^c\hat{h}_{lc}\right)} \\ \times e^{-\frac{1}{2}\ell''(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_1))\sum_c\left(\frac{1}{\sqrt{K}}\sum_l c_l\varphi'(\lambda_{l1})\,h_{lc}\right)^2 - \frac{1}{2}\ell'(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_1))\frac{1}{\sqrt{K}}\sum_{lc} c_l h_{lc}^2\varphi''(\lambda_{l1})} \end{aligned} \tag{89c}$$

where we have introduced the notation

$$y(\boldsymbol{\eta}) = f_\star\left[\frac{1}{\sqrt{K}}\sum_l c_l\,\varphi(\eta_l)\right] \tag{90a}$$

$$\hat{y}^\mu(\boldsymbol{\lambda}_a) = \frac{1}{\sqrt{K}}\sum_l c_l\,\varphi\left(\lambda_{la}\right) \tag{90b}$$

Note that the average over $\boldsymbol{w}^\star$ can be performed in the entropic term G_S. and we can simplify its expression to

$$G_S = \frac{1}{K}\log\int\prod_{la} dw_l^a\prod_{lic}\frac{du_{li}^c}{\sqrt{2\pi}} e^{\frac{1}{2}\sum_{abl}(\hat{q}_l^{ab}+\hat{r}_l^a\hat{r}_l^b)w_l^a w_l^b + \frac{1}{2}\sum_{cdl}(\hat{p}_l^{cd}+\hat{o}_l^c\hat{o}_l^d)u_l^c u_l^d + \sum_{acl}(\hat{t}_l^{ac}+\hat{r}_l^a\hat{o}_l^d)w_l^a u_l^c} \tag{91}$$

Similarly it is convenient to perform the integral over $\hat{\eta}_l$ in the energetic term

$$\begin{aligned} G_E = \log\int\prod_l D\eta_l\prod_{al}\frac{d\lambda_{la}d\hat{\lambda}_{la}}{2\pi}\prod_{cl}\frac{d\hat{h}_{lc}dh_{lc}}{2\pi} e^{i\sum_{la}(\lambda_{la}+r_l^a\eta_l)\hat{\lambda}_{la} + i\sum_{lc}(h_{lc}+o_l^c\eta_l)\hat{h}_{lc} - \beta\sum_a \ell(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_a))} \\ \times e^{-\frac{1}{2}\sum_l\left(\sum_{ab}(q_l^{ab}-r_l^a r_l^b)\hat{\lambda}_{la}\hat{\lambda}_{lb} + \sum_{cd}(p_l^{cd}-o_l^c o_l^d)\hat{h}_{lc}\hat{h}_{ld} + 2\sum_{ac}(t_l^{ac}-r_l^a o_l^c)\hat{\lambda}_{la}\hat{h}_{lc}\right)} \\ \times e^{-\frac{1}{2}\ell''(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_1))\sum_c\left(\frac{1}{\sqrt{K}}\sum_l c_l\varphi'(\lambda_{l1})\,h_{lc}\right)^2 - \frac{1}{2}\ell'(y(\boldsymbol{\eta}),\hat{y}(\boldsymbol{\lambda}_1))\frac{1}{\sqrt{K}}\sum_{lc} c_l h_{lc}^2\varphi''(\lambda_{l1})} \end{aligned} \tag{92}$$

where we remind the notation $D\eta \equiv \frac{d\eta}{\sqrt{2\pi}}e^{-\eta^2/2}$.

C.3 Annealed ansatz

In section B, we showed that the 1RSB theory provides an accurate approximation of the equilibrium measure, while higher-order RSB effects yield only negligible corrections. For simplicity, we therefore adopt a 1RSB ansatz for the overlap matrix q_l^{ab} (and the usual RS parametrization for r_l^a)

$$q_l^{ab} = q_0 + (q_1 - q_0)I_{ab}^{(n,m)} + (Q - q_1)I_{ab}^{(n,1)} \tag{93a}$$

$$\hat{q}_l^{ab} = \hat{q}_0 + (\hat{q}_1 - \hat{q}_0)I_{ab}^{(n,m)} + (\hat{Q} - \hat{q}_1)I_{ab}^{(n,1)} \tag{93b}$$

$$r_l^a = r \tag{93c}$$

$$\hat{r}_l^a = \hat{r} \tag{93d}$$

where we remind that $I_{ab}^{(n,m)}$ is the (a,b) element of a block matrix of size $n \times n$ whose diagonal blocks have size $m \times m$; those $m \times m$ blocks are filled with ones whereas outside them all the elements are zero. We do not have to worry of those order parameters because they fully satisfy the saddle point equations found in section B as in standard Franz-Parisi computations Franz &

Parisi (1995); Huang & Kabashima (2014). Moreover, the conjugated parameters can be analytically expressed in terms of the non-conjugated ones only.

We instead consider an annealed ansatz for the new overlaps

$$p_l^{cd} = P\delta_{cd} \tag{94a}$$

$$\hat{p}_l^{cd} = -\hat{P}\delta_{cd} \tag{94b}$$

$$t_l^{ac} = \hat{t}_l^{ac} = o_l^c = \hat{o}_l^c = 0 \tag{94c}$$

Inserting this ansatz one can easily perform the limit $n \to 0$ and $s \to 0$. The interaction and entropic term read

$$\mathcal{G}_I \equiv \lim_{\substack{n\to 0\\ s\to 0}} \partial_s G_I = \frac{1}{2}\hat{P}P + \frac{z}{2}P \tag{95a}$$

$$\mathcal{G}_S \equiv \lim_{\substack{n\to 0\\ s\to 0}} \partial_s G_S = -\frac{1}{2}\ln \hat{P} \tag{95b}$$

The energetic term is more involved, but it can be shown with standard methods to be equal to

$$\mathcal{G}_E = \lim_{\substack{n\to 0\\ s\to 0}} \partial_s G_E = \langle \log Z_h \rangle_\beta \ , \tag{96}$$

where Z_h and $\langle \bullet \rangle_\beta$ which represents the average over the equilibrium measure, are given by

$$\langle \bullet \rangle_\beta = \int D\boldsymbol{v}\, D\boldsymbol{x_1}\, \frac{\int D\boldsymbol{x_{01}}\, \mathcal{N}_0(\boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v})^{m-1} \int D\boldsymbol{x_0}\, e^{-\beta\, \ell(f_\star(\mathcal{O}_\star(\boldsymbol{v})),\, \mathcal{O}(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}))}\, \bullet}{\int D\boldsymbol{x_{01}}\, \mathcal{N}_0(\boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{\eta})^m} \tag{97a}$$

$$\mathcal{N}_0(\boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}) = \int D\boldsymbol{x_0}\, e^{-\beta\, \ell(f_\star(\mathcal{O}_\star(\boldsymbol{v})),\, \mathcal{O}(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}))} \tag{97b}$$

$$Z_h(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}) = \int D\boldsymbol{h}\, e^{-\frac{P}{2}\, \ell''(f_\star(\mathcal{O}_\star(\boldsymbol{v})),\, \mathcal{O}(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}))\, \mathcal{O}_1(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}, \boldsymbol{h})^2} \tag{97c}$$
$$\times\, e^{-\frac{P}{2}\, \ell'(f_\star(\mathcal{O}_\star(\boldsymbol{v})),\, \mathcal{O}(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}))\, \mathcal{O}_2(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}, \boldsymbol{h})}$$

We have also defined the following quantities

$$\mathcal{O}_\star(\boldsymbol{v}) = \frac{1}{\sqrt{K}} \sum_{l=1}^K c_l\, \varphi(v_l), \tag{98a}$$

$$u_l(x_0, x_1, x_{01}, v) = \sqrt{Q - q_1}\, x_{0,l} + \sqrt{q_1 - q_0}\, x_{01,l} + \sqrt{q_0 - r^2}\, x_{1,l} + r\, v_l \tag{98b}$$

$$\mathcal{O}(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}) = \frac{1}{\sqrt{K}} \sum_{l=1}^K c_l\, \varphi\left(u_l(x_0, x_1, x_{01}, v)\right) \tag{98c}$$

$$\mathcal{O}_1(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}, \boldsymbol{h}) = \frac{1}{\sqrt{K}} \sum_{l=1}^K c_l h_l\, \varphi'\left(u_l(x_0, x_1, x_{01}, v)\right) \tag{98d}$$

$$\mathcal{O}_2(\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_{01}}, \boldsymbol{v}, \boldsymbol{h}) = \frac{1}{\sqrt{K}} \sum_{l=1}^K c_l h_l^2\, \varphi''(u_l(x_0, x_1, x_{01}, v)) \ . \tag{98e}$$

Note that the expression of Z_h can be simplified by using Gaussian integration

$$\mathcal{G}_E = -\frac{1}{2} \left\langle \ln\det\left(I_K + P\mathcal{S}\right) \right\rangle_\beta \tag{99}$$

where we have introduced the $K \times K$ matrix $\mathcal{S}$ with elements

$$\mathcal{S}_{ll'} \equiv \frac{\ell''(f_\star(\mathcal{O}_\star), \mathcal{O})}{K} c_l c_{l'} \varphi'(u_l)\varphi'(u_{l'}) + \delta_{ll'} \frac{\ell'(f_\star(\mathcal{O}_\star), \mathcal{O})}{\sqrt{K}} c_l \varphi''(u_l) \tag{100}$$

Notice also that the saddle point equation for $\hat{P}$ can be readily solved as $\hat{P} = 1/P$. The entropic and interaction terms therefore simply read as follows

$$\mathcal{G}_{SI} \equiv \mathcal{G}_S + \mathcal{G}_I = \frac{1}{2} + \frac{1}{2}\ln P + \frac{z}{2}P \tag{101}$$

We therefore found that the "free entropy" $\phi(z)$ in (82) can be found by

$$\phi(z) = \max_P \left[\mathcal{G}_{SI} + \alpha \mathcal{G}_E\right] \tag{102}$$

The values of q, r and Q are found from the equilibrium saddle point equations (see section B). The maximization in (102) imposes the following saddle point equation for P:

$$z = -\frac{1}{P} + \alpha \left\langle \mathrm{Tr}\left[\mathcal{S}\left(I_K + P\mathcal{S}\right)^{-1}\right]\right\rangle_\beta \tag{103}$$

Note also that, following the discussion in Appendix C.1 the Stieltjes transform of the Hessian is equal to the value of $-P$ with P solving the saddle point equation (103), as

$$\mathcal{R}_{\alpha\mathcal{H}}(z) = -2\frac{\partial}{\partial z}\phi(z) = -P\,. \tag{104}$$

The spectrum is then computed using (75) which gives

$$\rho_{\alpha\mathcal{H}}(z) = -\frac{1}{\pi}\mathrm{Im}P\,. \tag{105}$$

C.4 Large width limit

Before sending K to infinity, we can furthermore simplify the energetic term, by using the matrix determinant lemma

$$\begin{aligned}\ln\det\left(I_K + P\mathcal{S}\right) = &\sum_l \log\left(1 + \frac{P\ell'(f_\star(\mathcal{O}_\star),\mathcal{O})}{\sqrt{K}}c_l\varphi''(u_l)\right)\\ &+ \log\left[1 + \frac{P\ell''(f_\star(\mathcal{O}_\star),\mathcal{O})}{K}\sum_l \frac{c_l^2\varphi'(u_l)^2}{1+\frac{P\ell'(f_\star(\mathcal{O}_\star),\mathcal{O})}{\sqrt{K}}c_l\varphi''(u_l)}\right]\,.\end{aligned} \tag{106}$$

When K is large we can Taylor expand this expression to get

$$\begin{aligned}\ln\det\left(I_K + P\mathcal{S}\right) \simeq &P\ell'(f_\star(\mathcal{O}_\star),\mathcal{O})\langle\mathcal{O}_2\rangle_h - \frac{P^2\ell'(f_\star(\mathcal{O}_\star),\mathcal{O})^2}{4}\left(\langle\mathcal{O}_2^2\rangle_h - \langle\mathcal{O}_2\rangle_h^2\right)\\ &+ \log\left[1 + P\ell''(f_\star(\mathcal{O}_\star),\mathcal{O})\langle\mathcal{O}_1^2\rangle_h\right]\end{aligned} \tag{107}$$

where we have used the identities

$$\langle\mathcal{O}_1^2\rangle_h = \frac{1}{K}\sum_l c_l^2\varphi'(u_l)^2 \tag{108a}$$

$$\langle\mathcal{O}_2\rangle_h = \frac{1}{\sqrt{K}}\sum_l c_l\varphi''(u_l) \tag{108b}$$

$$\langle\mathcal{O}_2^2\rangle_h - \langle\mathcal{O}_2\rangle_h^2 = \frac{2}{K}\sum_l c_l^2\varphi''(u_l)^2 \tag{108c}$$

In the large K limit the variables $\mathcal{O}_\star(\boldsymbol{v})$, $\mathcal{O}(\boldsymbol{x_0},\boldsymbol{x_1},\boldsymbol{x_{01}},\boldsymbol{\eta})$, $\mathcal{O}_1(\boldsymbol{x_0},\boldsymbol{x_1},\boldsymbol{x_{01}},\boldsymbol{v},\boldsymbol{h})$, $\mathcal{O}_2(\boldsymbol{x_0},\boldsymbol{x_1},\boldsymbol{x_{01}},\boldsymbol{v},\boldsymbol{h})$ in (98) become jointly Gaussian. The mean of all those observables vanishes because of the condition $\sum_l c_l = 0$. The covariance elements uniquely determined by the equilibrium overlaps Q, q_1, q_0, and r are those involving only averages of combinations of $\mathcal{O}$ and $\mathcal{O}\star$

$$\Phi_\star = \langle\mathcal{O}_\star^2\rangle_v - \langle\mathcal{O}_\star\rangle_v^2 = \mathcal{K}(1,1,1) - \mathcal{K}(1,1,0) \tag{109a}$$

$$D_0 = \langle\langle\mathcal{O}\rangle_{x_0,x_1,x_{01}}\mathcal{O}_\star\rangle_v - \langle\mathcal{O}\rangle\langle\mathcal{O}_\star\rangle = \mathcal{K}(1,Q,r) - \mathcal{K}(1,Q,0) \tag{109b}$$

$$\Phi_{Q,q_1} = \langle\mathcal{O}^2\rangle_{x_0,x_1,x_{01},v} - \langle\langle\mathcal{O}\rangle_{x_0}^2\rangle_{x_1,x_{01},v} = \mathcal{K}(Q,Q,Q) - \mathcal{K}(Q,Q,q_1) \tag{109c}$$

$$\Phi_{q_1,q_0} = \langle\langle\mathcal{O}\rangle_{x_0}^2\rangle_{x_1,x_{01},v} - \langle\langle\mathcal{O}\rangle_{x_0,x_{01}}^2\rangle_{x_1,v} = \mathcal{K}(Q,Q,q_1) - \mathcal{K}(Q,Q,q_0) \tag{109d}$$

$$\Phi_{q_0,0} = \langle\langle\mathcal{O}\rangle_{x_0,x_{01}}^2\rangle_{x_1,v} - \langle\mathcal{O}\rangle^2 = \mathcal{K}(Q,Q,q_0) - \mathcal{K}(Q,Q,0) \tag{109e}$$

which match the quantities introduced in Appendix B (see e.g. (36)) via the identification

$$\Phi_{a,b} = \Phi(a) - \Phi(b) \tag{110}$$

The new covariance elements involve the quantities $\mathcal{O}_1$ and $\mathcal{O}_2$ relevant for the determination of the Hessian. Those are

$$\Delta_0 \equiv \langle \mathcal{O}_1^2 \rangle \tag{111a}$$
$$\Xi_0 \equiv \langle \mathcal{O}_2^2 \rangle - \langle \langle \mathcal{O}_2 \rangle_h^2 \rangle_{x_0,x_1,x_{01},v} \tag{111b}$$
$$\Sigma_{Q,q_1} \equiv \langle \langle \mathcal{O}_2 \rangle_h \mathcal{O} \rangle_{x_0,x_1,x_{01},v} - \langle \langle \mathcal{O}_2 \rangle_{h,x_0} \langle \mathcal{O} \rangle_{x_0} \rangle_{x_1,x_{01},v} \tag{111c}$$
$$\Sigma_{q_1,q_0} \equiv \langle \langle \mathcal{O}_2 \rangle_{h,x_0} \langle \mathcal{O} \rangle_{x_0} \rangle_{x_1,x_{01},v} - \langle \langle \mathcal{O}_2 \rangle_{h,x_0,x_{01}} \langle \mathcal{O} \rangle_{x_0,x_{01}} \rangle_{x_1,v} \tag{111d}$$
$$\Sigma_{q_0,0} \equiv \langle \langle \mathcal{O}_2 \rangle_{h,x_0,x_{01}} \langle \mathcal{O} \rangle_{x_0,x_{01}} \rangle_{x_1,v} - \langle \mathcal{O}_2 \rangle \langle \mathcal{O} \rangle \tag{111e}$$
$$D_1 \equiv \langle \langle \mathcal{O}_2 \rangle_{h,x_0,x_1,x_{01}} \mathcal{O}_\star \rangle_v - \langle \mathcal{O}_2 \rangle \langle \mathcal{O}_\star \rangle \tag{111f}$$

Those effective order parameters can be all written in terms of the generalized kernel function (47) as

$$\Delta_0 = \mathcal{K}^{(1,1)}(Q,Q,Q) \tag{112a}$$
$$\Xi_0 = 2\,\mathcal{K}^{(2,2)}(Q,Q,Q) \tag{112b}$$
$$\Sigma_{Q,q_1} = \mathcal{K}^{(2,0)}(Q,Q,Q) - \mathcal{K}^{(2,0)}(Q,Q,q_1) \tag{112c}$$
$$\Sigma_{q_1,q_0} = \mathcal{K}^{(2,0)}(Q,Q,q_1) - \mathcal{K}^{(2,0)}(Q,Q,q_0) \tag{112d}$$
$$\Sigma_{q_0,0} = \mathcal{K}^{(2,0)}(Q,Q,q_0) - \mathcal{K}^{(2,0)}(Q,Q,0) \tag{112e}$$
$$D_1 = \mathcal{K}^{(0,2)}(1,Q,r) - \mathcal{K}^{(0,2)}(1,Q,0) \tag{112f}$$

Using those covariances one finally finds that the energetic term has the form

$$\mathcal{G}_E = \frac{1}{2}A_2 P^2 + \frac{1}{2}A_1 P - \frac{1}{2}\left\langle \ln\left(1 + P\Delta_0 \ell''(y,\hat{y})\right)\right\rangle_\beta \tag{113}$$

with the coefficients A_2 and A_1 being independent of P

$$A_2 = \frac{\Xi_0}{4}\left\langle \ell'(y,\hat{y})^2 \right\rangle_\beta \tag{114a}$$
$$A_1 = -\left\langle \ell'(y,\hat{y}) \left(\frac{D_1}{\sqrt{\Phi_\star}} v + \frac{\Sigma_{Q,q_1}}{\sqrt{\Phi_{Q,q_1}}} x_0 + \frac{\Sigma_{q_0,0} - \frac{D_0 D_1}{\Phi_\star}}{\sqrt{\Phi_{q_0,0} - \frac{D_0^2}{\Phi_\star}}} x_1 + \frac{\Sigma_{q_1,q_0}}{\sqrt{\Phi_{q_1,q_0}}} x_{01} \right) \right\rangle_\beta \tag{114b}$$

In the previous expression we have denoted by $\langle \bullet \rangle_\beta$ the energetic equilibrium measure average in the large K limit, which at 1RSB level reads

$$\langle \bullet \rangle_\beta \equiv \int Dv Dx_1 \frac{\int Dx_{01} \left[\int Dx_0\, e^{-\beta \ell(y,\hat{y})}\right]^m \frac{\int Dx_0\, e^{-\beta\ell(y,\hat{y})} \bullet}{\int Dx_0\, e^{-\beta\ell(y,\hat{y})}}}{\int Dx_{01} \left[\int Dx_0\, e^{-\beta\ell(y,\hat{y})}\right]^m} \tag{115a}$$
$$y \equiv f_\star\left(\sqrt{\Phi_\star} v\right), \tag{115b}$$
$$\hat{y} \equiv \frac{D_0}{\sqrt{\Phi_\star}} v + \sqrt{\Phi_{Q,q_1}} x_0 + \sqrt{\Phi_{q_0,0} - \frac{D_0^2}{\Phi_\star}} x_1 + \sqrt{\Phi_{q_1,q_0}} x_{01}\,. \tag{115c}$$

C.5 Square loss case

When one takes the square loss $\ell(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2$ all the integrals in the energetic term are Gaussian and can be solved. The free entropy therefore is given by $\phi = \mathcal{G}_{SI} + \alpha \mathcal{G}_E$ where

$$\mathcal{G}_{SI} = \frac{1}{2} + \frac{z}{2}P + \frac{1}{2}\ln P \tag{116a}$$
$$\mathcal{G}_E = \frac{1}{2}A_2 P^2 + \frac{1}{2}A_1 P - \frac{1}{2}\ln(1 + P\Delta_0) \tag{116b}$$

with the coefficients A_2 and A_1 that read

$$A_2 = \frac{\Xi_0}{4(1+\beta\Phi_{Q,q_1})^2}\left[\Phi_{Q,q_1}(1+\beta\Phi_{Q,q_1}) + \frac{\Phi_{q_1,q_0}}{B} + \frac{C}{B^2}\right] \tag{117a}$$

$$A_1 = -\frac{1}{(1+\beta\Phi_{Q,q_1})^2}\left[(1+\beta\Phi_{Q,q_1})\left(\Sigma_{Q,q_1}+\frac{T}{B}\right)-\frac{\beta\Sigma_{Q,q_1}}{B}\left(\Phi_{q_1,q_0}+\frac{C}{B}\right)-\Sigma_{q_1,q_0}\frac{m\beta}{B^2}C\right] \tag{117b}$$

$$B = 1+\frac{m\beta\Phi_{q_1,q_0}}{1+\beta\Phi_{Q,q_1}} \tag{117c}$$

The constants C and T depend on the task, i.e. they differ from the classification ($f_\star(\hat{y}) = \mathrm{sign}(\hat{y})$) to the regression case ($f_\star(\hat{y}) = \hat{y}$). In classification one has

$$C_{\text{class}} = 1+\Phi_{q_0,0}-2D_0\sqrt{\frac{2}{\pi\Phi_\star}} \tag{118a}$$

$$T_{\text{class}} = \Sigma_{q_0,0}+\Sigma_{q_1,q_0}-D_1\sqrt{\frac{2}{\pi\Phi_\star}} \tag{118b}$$

whereas in regression

$$C_{\text{reg}} = \Phi_\star+\Phi_{q_0,0}-2D_0 \tag{119a}$$

$$T_{\text{reg}} = \Sigma_{q_0,0}+\Sigma_{q_1,q_0}-D_1 \tag{119b}$$

The saddle point equation for P in both cases is therefore a cubic equation

$$1+\tilde{a}_1(z)P+\tilde{a}_2(z)P^2+\tilde{a}_3P^3 = 0 \tag{120}$$

with coefficients given by

$$\tilde{a}_1(z) = z+(1-\alpha)\Delta_0+\alpha A_1 \tag{121a}$$

$$\tilde{a}_2(z) = z\Delta_0+\alpha(2A_2+A_1\Delta_0) \tag{121b}$$

$$\tilde{a}_3 = 2\alpha\Delta_0A_2\,. \tag{121c}$$

Note that the coefficients in the RS case can be obtained by setting $q_1 = q_0 = q$ that is $\Phi_{q_1,q_0} = \Sigma_{q_1,q_0} = 0$.

Equation (120) can be therefore solved exactly. Indeed the cubic equation has either three real roots, corresponding to a vanishing spectral density, or one real root and a complex-conjugate pair. In the latter case, because of (105) the physical solution is the root with negative imaginary part, so that $\rho_{\alpha\mathcal{H}}(z) = -(1/\pi)\,\mathrm{Im}\,P(z) \geq 0$. Finally, recalling that the computation above concerns $\alpha\mathcal{H}$, we have $\mathcal{R}_{\alpha\mathcal{H}}(z) = -P(z)$ and $\mathcal{R}_{\mathcal{H}}(z) = \alpha\mathcal{R}_{\alpha\mathcal{H}}(\alpha z)$. Therefore (120) evaluated at αz, yields equation (16) of the main text with the identification

$$a_1(z) = -\frac{\widetilde{a}_1(\alpha z)}{\alpha},\qquad a_2(z) = \frac{\widetilde{a}_2(\alpha z)}{\alpha^2},\qquad a_3 = -\frac{\widetilde{a}_3}{\alpha^3}. \tag{122}$$

C.6 Zero temperature limit

Remind that in the large β limit we have to impose the following scalings

$$q_1 = Q-\frac{\delta q}{\beta} \tag{123a}$$

$$m = \frac{\delta m}{\beta} \tag{123b}$$

This implies the following scalings in β for the effective order parameters

$$\Phi_{Q,q_1} = \frac{\delta\Phi}{\beta} \tag{124a}$$

$$\Sigma_{Q,q_1} = \frac{\delta\Sigma}{\beta} \tag{124b}$$

It is easy to show by Taylor expansion that $\delta\Phi$ and $\delta\Sigma$ are expressed in terms of the generalized kernel function in (47) as

$$\delta\Phi = \mathcal{K}^{(1,1)}(Q,Q,Q)\,\delta q \tag{125a}$$

$$\delta\Sigma = \mathcal{K}^{(3,1)}(Q,Q,Q)\,\delta q \tag{125b}$$

Note that the other effective order parameters stay finite in the large β limit. One finally finds

$$\lim_{\beta\to\infty}\mathcal{G}_E = -\frac{P}{2}\left\langle \ell'(y,\delta\hat{y}_\star)\left(\frac{D_1}{\sqrt{\Phi_\star}}v + \frac{\delta\Sigma}{\sqrt{\delta\Phi}}x_0^\star(v,x_{01},x_1) + \frac{\Sigma_{q_0,0}-\frac{D_0D_1}{\Phi_\star}}{\sqrt{\Phi_{q_0,0}-\frac{D_0^2}{\Phi_\star}}}x_1 + \frac{\Sigma_{Q,q_0}}{\sqrt{\Phi_{Q,q_0}}}x_{01}\right)\right\rangle_\infty$$
$$+\frac{\Xi_0 P^2}{8}\left\langle \ell'(y,\delta\hat{y}_\star)^2\right\rangle_\infty - \frac{1}{2}\left\langle \ln\left(1+P\Delta_0\ell''(y,\delta\hat{y}_\star)\right)\right\rangle_\infty \tag{126}$$

where

$$x_0^\star(v,x_{01},x_1) \equiv \underset{x_0}{\operatorname{argmax}}\left[-\frac{x_0^2}{2} - \ell\left(y(v),\delta y(v,x_0,x_{01},x_1)\right)\right] \tag{127a}$$

$$\delta\hat{y}(v,x_0,x_1,x_{01}) \equiv \frac{D_0}{\sqrt{\Phi_\star}}v + \sqrt{\delta\Phi}x_0 + \sqrt{\Phi_{q_0,0}-\frac{D_0^2}{\Phi_\star}}x_1 + \sqrt{\Phi_{Q,q_0}}x_{01} \tag{127b}$$

$$y(v) \equiv f_\star\left(\sqrt{\Phi_\star}v\right) \tag{127c}$$

$$\delta\hat{y}_\star = \delta\hat{y}(v,x_0^\star(v,x_{01},x_1),x_1,x_{01}) \tag{127d}$$

with the equilibrium measure in the large β limit being

$$\langle\bullet\rangle_\infty \equiv \int DvDx_1\,\frac{\int Dx_{01}\,e^{-\delta m\,\ell(y,\delta\hat{y}_\star)}\,\bullet}{\int Dx_{01}\,e^{-\delta m\,\ell(y,\delta\hat{y}_\star)}} \tag{128}$$

The replica symmetric case can be again recovered setting $\Phi_{Q,q_0} = \Sigma_{Q,q_0} = 0$.

C.6.1 Square loss case

In the square loss case the limit can be worked out directly from the expressions given in section C.5. What will change is the expression of the coefficients A_1 and A_2. They are given by

$$A_2 = \frac{\Xi_0}{4(1+\delta\Phi)^2}\left[\frac{\Phi_{Q,q_0}}{B} + \frac{C}{B^2}\right] \tag{129a}$$

$$A_1 = -\frac{1}{(1+\delta\Phi)^2}\left[(1+\delta\Phi)\frac{T}{B} - \frac{\delta\Sigma}{B}\left(\Phi_{Q,q_0} + \frac{C}{B}\right) - \Sigma_{Q,q_0}\,\frac{\delta m}{B^2}\,C\right] \tag{129b}$$

$$B = 1 + \frac{\delta m\Phi_{Q,q_0}}{1+\delta\Phi} \tag{129c}$$

where we remind that the constants C and T depend on the task and are given by (118) and (119) respectively in classification and regression, and tend to a finite limit for large β.

C.7 Left and right edges of the spectrum of the Hessian

We remind the reader that the Stieltjes transform of the Hessian and the spectrum are respectively given by (see Appendix C.1)

$$\mathcal{R}_{\alpha\mathcal{H}}(z) = -2\frac{\partial}{\partial z}\phi(z) = -P \tag{130a}$$

$$\rho_{\alpha\mathcal{H}}(z) = -\frac{1}{\pi}\mathrm{Im}P\,, \tag{130b}$$

and the value of P is found maximizing $\phi(z)$ with respect to P, which gives

$$z = -\frac{1}{P} + \alpha\left[-2A_2P - A_1 + \left\langle\frac{\Delta_0\ell''(y,\hat{y})}{1+P\Delta_0\ell''(y,\hat{y})}\right\rangle_\beta\right] \tag{131}$$

The edges of the spectrum can be obtained by looking for the points where the derivative of $z(P)$ is equal to zero, as beyond these points $z(P)$ is not defined for real P. Setting $\frac{dz}{dP} = 0$ we get the two solutions (corresponding to the left and right edges)

$$P_\pm = \pm\frac{1}{\sqrt{\alpha}}\left[2A_2 + \left\langle\left(\frac{\Delta_0\ell''(y,\hat{y})}{1+P_\pm\Delta_0\ell''(y,\hat{y})}\right)^2\right\rangle_\beta\right]^{-1/2}. \tag{132}$$

C.8 A COMMENT ON THE VALIDITY OF THE ANNEALED ANSATZ

It is possible to generalize the previous computation which was valid for an annealed structure of the order parameters (94), to a replica symmetric one

$$p_l^{cd} = P\delta_{cd} + (1-\delta_{cd})p \tag{133a}$$
$$\hat{p}_l^{cd} = -\hat{P}\delta_{cd} + (1-\delta_{cd})\hat{p} \tag{133b}$$
$$t_l^{ac} = T\delta_{a1} + (1-\delta_{a1})t \tag{133c}$$
$$\hat{t}_l^{ac} = \hat{T}\delta_{a1} + (1-\delta_{a1})\hat{t} \tag{133d}$$
$$o_l^{c} = o \tag{133e}$$
$$\hat{o}_l^{c} = \hat{o} \tag{133f}$$

The computation is however long and tedious (especially when dealing with the large width limit) and we do not report it here. However one can analyze straightforwardly the saddle point equations. First the equation for the conjugated parameters $\hat{P}$, $\hat{T}$, $\hat{t}$ and $\hat{o}$ can be solved explicitly in terms of P, p, T, t, o. One finds then that the free entropy function is a quadratic form of the variables T, t and o. The only solution of the saddle point equations is therefore $T = t = o = 0$. Inserting this solution inside the free entropy one also finds that $p = 0$ is always a solution. In this way one finds that

$$t = \hat{t} = T = \hat{T} = p = \hat{p} = 0\,. \tag{134}$$

This tells us that the annealed ansatz is the correct one.

D DETAILS OF THE NUMERICAL EXPERIMENTS

We validate the theory with full-batch gradient-descent simulations of the model of Section 2, using $N = 2000$, $K = 20$, and $\varphi = \mathrm{erf}$. For each α we draw a teacher 20 times, generate $P = \lfloor \alpha N \rfloor$ i.i.d. standard Gaussian training inputs and an independent test set of size 10000, and minimize the regularized objective (3) with nominal weight decay λ and learning rate $\eta_0 \in [0.5, 1.0]$. Training runs for up to $E_{\max} \in [10^4, 10^5]$ epochs, with early stopping when the relative change of the loss and of the order parameters $Q = \|\boldsymbol{w}\|^2/N$ and $r = \boldsymbol{w}\cdot\boldsymbol{w}^\star/N$ all fall below $\varepsilon = 10^{-8}$ between two consecutive checkpoints (checked every 100 epochs). At the end of training we evaluate the Hessian of the unregularized loss in closed form, exploiting the committee structure, and diagonalize the full $N \times N$ matrix exactly, without any low-rank or stochastic approximation. The experiments are implemented in Julia, using Flux.jl for training, ForwardDiff.jl for φ', φ'', and LinearAlgebra.jl for the diagonalization.