The committee machine: Computational to statistical gaps in learning a two-layers neural network

Benjamin Aubin[⋆†], Antoine Maillard[†], Jean Barbier[◊],
Florent Krzakala[†], Nicolas Macris[⊗] and Lenka Zdeborová[⋆]

Abstract

Heuristic tools from statistical physics have been used in the past to locate the phase transitions and compute the optimal learning and generalization errors in the teacher-student scenario in multi-layer neural networks. In this contribution, we provide a rigorous justification of these approaches for a two-layers neural network model called the committee machine, under a technical assumption. We also introduce a version of the approximate message passing (AMP) algorithm for the committee machine that allows to perform optimal learning in polynomial time for a large set of parameters. We find that there are regimes in which a low generalization error is information-theoretically achievable while the AMP algorithm fails to deliver it; strongly suggesting that no efficient algorithm exists for those cases, and unveiling a large computational gap.

Contents

⋆ Institut de Physique Théorique, CNRS & CEA & Université Paris-Saclay, Saclay, France.
† Laboratoire de Physique Statistique, CNRS & Sorbonnes Universités & École Normale Supérieure, PSL University, Paris, France.
⊗ Laboratoire de Théorie des Communications, École Polytechnique Fédérale de Lausanne, Suisse.
◊ International Center for Theoretical Physics, Trieste, Italy.

1 Introduction

While the traditional approach to learning and generalization follows the Vapnik-Chervonenkis [1] and Rademacher [2] worst-case type bounds, there has been a considerable body of theoretical work on calculating the generalization ability of neural networks for data arising from a probabilistic model within the framework of statistical mechanics [3, 4, 5, 6, 7]. In the wake of the need to understand the effectiveness of neural networks and also the limitations of the classical approaches [8], it is of interest to revisit the results that have emerged thanks to the physics perspective. This direction is currently experiencing a strong revival, see e.g. [9, 10, 11, 12].

Of particular interest is the so-called teacher-student approach, where labels are generated by feeding i.i.d. random samples to a neural network architecture (the *teacher*) and are then presented to another neural network (the *student*) that is trained using these data. Early studies computed the information theoretic limitations of the supervised learning abilities of the teacher weights by a student who is given m independent n-dimensional examples with $\alpha \equiv m/n = \Theta(1)$ and $n \to \infty$ [3, 4, 7]. These works relied on non-rigorous heuristic approaches, such as the replica and cavity methods [13, 14]. Additionally no provably efficient algorithm was provided to achieve the predicted learning abilities, and it was thus difficult to test those predictions, or to assess the computational difficulty.

Recent developments in statistical estimation and information theory —in particular of approximate message passing algorithms (AMP) [15, 16, 17, 18], and a rigorous proof of the replica formula for the optimal generalization error [11]— allowed to settle these two missing points for single-layer neural networks (i.e. without any hidden variables). In the present paper, we leverage on these works, and provide rigorous asymptotic predictions and corresponding message passing algorithm for a class of two-layers networks.

2 Summary of contributions and related works

While our results hold for a rather large class of non-linear activation functions, we illustrate our findings on a case considered most commonly in the early literature: the committee machine. This is possibly the simplest version of a two-layers neural network where all the weights in the second layer are fixed to unity, and we illustrate it in Fig. 1. Denoting Y_μ the label associated with a n-dimensional sample X_μ, and W^*_{il} the weight connecting the i-th coordinate of the input to the l-th node of the hidden layer, it is defined by:

$$Y_\mu = \text{sign}\Big[\sum_{l=1}^{K} \text{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W^*_{il}\Big)\Big] . \tag{1}$$

We concentrate here on the teacher-student scenario: The teacher generates i.i.d. data samples with i.i.d. standard Gaussian coordinates $X_{\mu i} \sim \mathcal{N}(0,1)$, then she/he generates the associated labels Y_μ using a committee machine as in (1), with i.i.d. weights W^*_{il} unknown to the student (in the proof section we will consider the more general case of a distribution for the weights of the form $\prod_{i=1}^n P_0(\{W^*_{il}\}_{l=1}^K)$, but in practice we consider the fully separable case). The student is then given the m input-output pairs $(X_\mu, Y_\mu)_{\mu=1}^m$ and knows the distribution P_0 used to generate W^*_{il}. The goal of the student is to learn the weights W^*_{il} from the available examples $(X_\mu, Y_\mu)_{\mu=1}^m$ in order to reach the smallest possible generalization error (i.e. to be able to predict the label the teacher would generate for a new sample not present in the training set).

There have been several studies of this model within the non-rigorous statistical physics approach in the limit where $\alpha \equiv m/n = \Theta(1)$, $K = \Theta(1)$ and $n \to \infty$ [19, 20, 21, 22, 6, 7]. A particularly interesting result in the teacher-student setting is the *specialization of hidden neurons* (see sec. 12.6 of [7], or [23] in the context of online learning): For $\alpha < \alpha_{\text{spec}}$ (where α_{spec} is a certain critical value of the sample complexity),

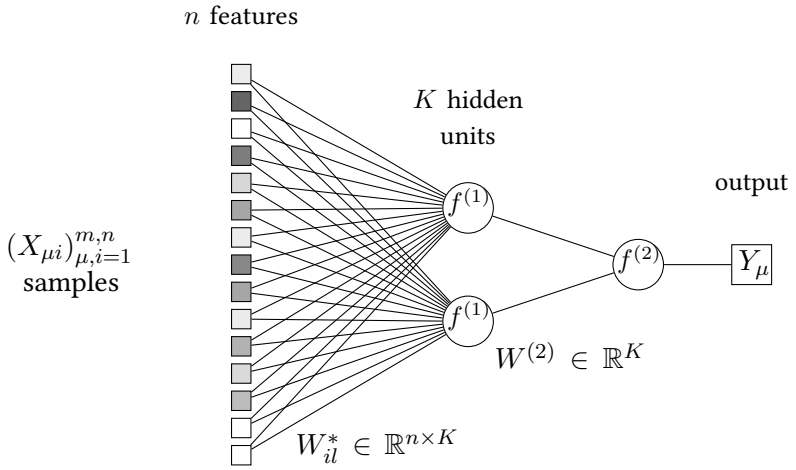

Figure 1: The *committee machine* is one of the simplest models belonging to the considered model class (2), and on which we focus to illustrate our results. It is a two-layers neural network with activation sign functions $f^{(1)}, f^{(2)} =$ sign and weights $W^{(2)}$ fixed to unity. It is represented for $K = 2$.

the permutational symmetry between hidden neurons remains conserved even after an optimal learning, and the learned weights of each of the hidden neurons are identical. For $\alpha > \alpha_{\rm spec}$, however, this symmetry gets broken as each of the hidden units correlates strongly with one of the hidden units of the teacher. Another remarkable result is the calculation of the optimal generalization error as a function of α.

Our first contribution consists in a proof of the replica formula conjectured in the statistical physics literature, using the adaptive interpolation method of [24, 11], that allows to put several of these results on a rigorous basis. This proof uses a technical unproven assumption. Our second contribution is the design of an AMP-type of algorithm that is able to achieve the optimal generalization error in the above limit of large dimensions for a wide range of parameters. The study of AMP —that is widely believed to be optimal between all polynomial algorithms in the above setting [25, 26, 27, 28]— unveils, in the case of the committee machine with a large number of hidden neurons, the existence a large *hard phase* in which learning is information-theoretically possible, leading to a good generalization error decaying asymptotically as $1.25K/\alpha$ (in the $\alpha = \Theta(K)$ regime), but where AMP fails and provides only a poor generalization that does not go to zero when increasing α. This strongly suggests that no efficient algorithm exists in this hard region and therefore there is a computational gap in learning such neural networks. In other problems where a hard phase was identified its study boosted the development of algorithms that are able to match the predicted thresholds and we anticipate this will translate to the present model.

We also want to comment on a related line of work that studies the loss-function landscape of neural networks. While a range of works show under various assumptions that spurious local minima are absent in neural networks, others show under different conditions that they do exist, see e.g. [29]. The regime of parameters that is hard for AMP must have spurious local minima, but the converse is not true in general. It might be that there are spurious local minima, yet the AMP approach succeeds. Moreover, in all previously studied models in the Bayes-optimal setting the (generalization) error obtained with the AMP is the best known and other approaches, e.g. (noisy) gradient based, spectral algorithms or semidefinite programming, are not better in generalizing even in cases where the "student" models are overparametrized. Of course in order to be in the Bayes-optimal setting one needs to know the model used by the teacher which is not the case in practice.

3 Main technical results

3.1 A general model

While in the illustration of our results we shall focus on the model (1), all our formulas are valid for a broader class of models: Given m input samples $(X_{\mu i})_{\mu,i=1}^{m,n}$, we denote W_{il}^* the teacher-weight connecting the i-th input (i.e. visible unit) to the l-th node of the hidden layer. For a generic function $\varphi_{\rm out} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ one can formally write the output as

$$Y_\mu = \varphi_{\rm out}\Big(\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}W_{il}^*\Big\}_{l=1}^K, A_\mu\Big) \quad \text{or} \quad Y_\mu \sim P_{\rm out}\Big(\cdot\,\Big|\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}W_{il}^*\Big\}_{l=1}^K\Big)\,, \tag{2}$$

where $(A_\mu)_{\mu=1}^m$ are i.i.d. real valued random variables with known distribution P_A, that form the probabilistic part of the model, generally accounting for noise.

For deterministic models the second argument is simply absent (or is a Dirac mass). We can view alternatively (2) as a channel where the transition kernel $P_{\rm out}$ is directly related to $\varphi_{\rm out}$. As discussed above, we focus on the teacher-student scenario where the teacher generates Gaussian i.i.d. data $X_{\mu i} \sim \mathcal{N}(0,1)$, and i.i.d. weights $W_{il}^* \sim P_0$. The student then learns W^* from the data $(X_\mu, Y_\mu)_{\mu=1}^m$ by computing marginal means of the posterior probability distribution (5).

Different scenarii fit into this general framework. Among those, the committee machine is obtained when choosing $\varphi_{\rm out}(h) = \text{ sign}(\sum_{l=1}^K \text{ sign}(h_l))$ while another model considered previously is given by the parity machine, when $\varphi_{\rm out}(h) = \prod_{l=1}^K \text{ sign}(h_l)$, see e.g. [7] and sec. H for the numerical results in the case $K = 2$. A number of layers beyond two has also been considered, see [22]. Other activation functions can be used, and many more problems can be described, e.g. compressed pooling [30, 31] or multi-vector compressed sensing [32].

3.2 Two auxiliary inference problems

Denote $\mathcal{S}_K$ the finite dimensional vector space of $K \times K$ matrices, $\mathcal{S}_K^+$ the convex set of semi-definite positive $K \times K$ matrices, $\mathcal{S}_K^{++}$ for positive definite $K \times K$ matrices, and $\forall\, N \in \mathcal{S}_K^+$ we set $S_K^+(N) \equiv \{M \in S_K^+ \text{ s.t. } N - M \in \mathcal{S}_K^+\}$. Note that $\mathcal{S}_K^+(N)$ is convex and compact.

Stating our results requires introducing two simpler auxiliary K-dimensional estimation problems:

• The first one consists in retrieving a K-dimensional input vector $W_0 \sim P_0$ from the output of a Gaussian vector channel with K-dimensional observations

$$Y_0 = r^{1/2}W_0 + Z_0\,,$$

$Z_0 \sim \mathcal{N}(0, I_{K\times K})$ and the "channel gain" matrix $r \in \mathcal{S}_K^+$. The posterior distribution on $w = (w_l)_{l=1}^K$ is

$$P(w|Y_0) = \frac{1}{\mathcal{Z}_{P_0}} P_0(w) e^{Y_0^\intercal r^{1/2} w - \frac{1}{2}w^\intercal r w}\,, \tag{3}$$

and the associated *free entropy* (or minus *free energy*) is given by the expectation over Y_0 of the log-partition function

$$\psi_{P_0}(r) \equiv \mathbb{E}\ln \mathcal{Z}_{P_0}$$

and involves K dimensional integrals.

• The second problem considers K-dimensional i.i.d. vectors $V, U^* \sim \mathcal{N}(0, I_{K\times K})$ where V is considered to

be known and one has to retrieve U^* from a scalar observation obtained as

$$\widetilde{Y}_0 \sim P_{\text{out}}(\,\cdot\,|q^{1/2}V + (\rho - q)^{1/2}U^*)$$

where the second moment matrix $\rho \equiv \mathbb{E}[W_0 W_0^\intercal]$ is in $\mathcal{S}_K^+$ (where $W_0 \sim P_0$) and the so-called "overlap matrix" q is in $S_K^+(\rho)$. The associated posterior is

$$P(u|\widetilde{Y}_0, V) = \frac{1}{\mathcal{Z}_{P_{\text{out}}}} \frac{e^{-\frac{1}{2}u^\intercal u}}{(2\pi)^{K/2}} P_{\text{out}}\big(\widetilde{Y}_0|q^{1/2}V + (\rho - q)^{1/2}u\big)\,, \tag{4}$$

and the free entropy reads this time

$$\Psi_{P_{\text{out}}}(q;\rho) \equiv \mathbb{E}\ln \mathcal{Z}_{P_{\text{out}}}$$

(with the expectation over $\widetilde{Y}_0$ and V) and also involves K dimensional integrals.

3.3 The free entropy

The central object of study leading to the optimal learning and generalization errors in the present setting is the posterior distribution of the weights:

$$P(\{w_{il}\}_{i,l=1}^{n,K} \mid \{X_{\mu i}, Y_\mu\}_{\mu,i=1}^{m,n}) = \frac{1}{\mathcal{Z}_n} \prod_{i=1}^{n} P_0(\{w_{il}\}_{l=1}^{K}) \prod_{\mu=1}^{m} P_{\text{out}}\Big(Y_\mu\Big|\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{\mu i} w_{il}\Big\}_{l=1}^{K}\Big)\,, \tag{5}$$

where the normalization factor is nothing else than a *partition function*, i.e. the integral of the numerator over $\{w_{il}\}_{i,l=1}^{n,K}$. The expected[1] free entropy is by definition

$$f_n \equiv \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_n\,. \tag{6}$$

The replica formula gives an explicit (conjectural) expression of f_n in the high-dimensional limit $n, m \to \infty$ with $\alpha = m/n$ fixed. We show in sec. B how the heuristic replica method [13, 14] yields the formula. This computation was first performed, to the best of our knowledge, by [19] in the case of the committee machine. Our first contribution is a rigorous proof of the corresponding free entropy formula using an interpolation method [33, 34, 24], under a technical Assumption 1.

In order to formulate our results, we add an (arbitrarily small) Gaussian regularization noise $Z_\mu\sqrt{\Delta}$ to the first expression of the model (2), where $\Delta > 0$, $Z_\mu \sim \mathcal{N}(0,1)$, which thus becomes

$$Y_\mu = \varphi_{\text{out}}\Big(\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{\mu i} W_{il}^*\Big\}_{l=1}^{K}, A_\mu\Big) + Z_\mu\sqrt{\Delta}\,, \tag{7}$$

so that the channel kernel is ($u \in \mathbb{R}^K$)

$$P_{\text{out}}(y|u) = \frac{1}{\sqrt{2\pi\Delta}} \int_{\mathbb{R}} dP_A(a) e^{-\frac{1}{2\Delta}(y - \varphi_{\text{out}}(u,a))^2}\,. \tag{8}$$

Let us define the *replica symmetric (RS) potential* as

$$f_{\text{RS}}(q,r) = f_{\text{RS}}(q,r;\rho) \equiv \psi_{P_0}(r) + \alpha\Psi_{P_{\text{out}}}(q;\rho) - \frac{1}{2}\text{Tr}(rq), \tag{9}$$

[1]The symbol $\mathbb{E}$ will generally denote an expectation over all random variables in the ensuing expression (here $\{X_{\mu i}, Y_\mu\}$). Subscripts will be used only when we take partial expectations or if there is an ambiguity.

where $\alpha \equiv m/n$, and $\Psi_{P_{\text{out}}}(q;\rho)$ and $\psi_{P_0}(r)$ are the free entropies of the two simpler K-dimensional estimation problems (3) and (4).

All along this paper, we assume the following hypotheses for our rigorous statements:

(H1) The prior P_0 has bounded support in $\mathbb{R}^K$.
(H2) The activation $\varphi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ is a bounded $\mathcal{C}^2$ function with bounded first and second derivatives w.r.t. its first argument (in $\mathbb{R}^K$-space).
(H3) For all $\mu = 1, \dots, m$ and $i = 1, \dots, n$ we have i.i.d. $X_{\mu i} \sim \mathcal{N}(0,1)$.

We finally rely on a technical hypothesis, stated as Assumption 1 in section 5.3.

Theorem 3.1 (Replica formula). *Suppose (H1), (H2) and (H3), and Assumption 1. Then for the model (7) with kernel (8) the limit of the free entropy is:*

$$\lim_{n\to\infty} f_n \equiv \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_n = \sup_{r\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\text{RS}}(q,r)\,. \tag{10}$$

This theorem extends the recent progress for generalized linear models of [11], which includes the case $K = 1$ of the present contribution, to the phenomenologically richer case of two-layers problems such as the committee machine. The proof sketch based on an *adaptive interpolation method* recently developed in [24] is outlined in sec. 5 and the details can be found in sec. A.

Remark 3.2 (Relaxing the hypotheses). *Note that, following similar approximation arguments as in [11], the hypothesis (H1) can be relaxed to the existence of the second moment of the prior; thus covering the Gaussian case, (H2) can be dropped (and thus include model (1) and its sign$(\cdot)$ activation) and (H3) extended to data matrices X with i.i.d. entries of zero mean, unit variance and finite third moment. Moreover, the case $\Delta = 0$ can be considered when the outputs are discrete, as in the committee machine (1), see [11]. The channel kernel becomes in this case $P_{\text{out}}(y|u) = \int dP_A(a)\mathbf{1}(y - \varphi_{\text{out}}(u,a))$ and the replica formula is the limit $\Delta \to 0$ of the one provided in Theorem 3.1. In general this regularizing noise is needed for the free entropy limit to exist.*

3.4 Learning the teacher weights and optimal generalization error

A classical result in Bayesian estimation is that the estimator $\hat{W}$ that minimizes the mean-square error with the ground-truth W^* is given by the expected mean of the posterior distribution. Denoting q^* the extremizer in the replica formula (10), we expect from the replica method that in the limit $n \to \infty, m/n = \alpha$, and with high probablity, $\hat{W}^\intercal W^*/n \to q^*$. We refer to proposition 5.3 and to the proof in sec. A for the precise statement, that remains rigorously valid *only* in the presence of an additional (possibly infinitesimal) side-information. From the overlap matrix q^*, one can compute the Bayes-optimal generalization error when the student tries to classify a new, yet unseen, sample X_{new}. The estimator of the new label $\hat{Y}_{\text{new}}$ that minimizes the mean-square error with the true label is given by computing the posterior mean of $\varphi_{\text{out}}(X_{\text{new}}w)$ (X_{new} is a row vector). Given the new sample, the optimal generalization error is then

$$\frac{1}{2}\mathbb{E}_{X,W^*}\left[\left(\mathbb{E}_{w|X,Y}\left[\varphi_{\text{out}}(X_{\text{new}}w)\right] - \varphi_{\text{out}}(X_{\text{new}}W^*)\right)^2\right] \xrightarrow[n\to\infty]{} \epsilon_g(q^*), \tag{11}$$

where w is distributed according to the posterior measure (5) (note that this Bayes-optimal computation differs from the so-called Gibbs estimator by a factor 2, see sec. C). In particular, when the data X is drawn from the standard Gaussian distribution on $\mathbb{R}^{m\times n}$, and is thus rotationally invariant, it follows that this error only depends on $w^\intercal W^*/n$, which converges to q^*. Then a direct algebraic computation gives a lengthy but explicit formula for $\epsilon_g(q^*)$, as shown in sec. C.

3.5 Approximate message passing, and its state evolution

Our next result is based on an adaptation of a popular algorithm to solve random instances of generalized linear models, the *Approximate Message Passing* (AMP) algorithm [15, 16], for the case of the committee machine and models described by (2).

The AMP algorithm can be obtained as a Taylor expansion of loopy belief-propagation (see sec. F) and also originates in earlier statistical physics works [35, 36, 37, 38, 39, 26]. It is conjectured to perform the best among all polynomial algorithms in the framework of these models. It thus gives us a tool to evaluate both the intrinsic algorithmic hardness of the learning and the performance of existing algorithms with respect to the optimal one in this model.

Algorithm 1 Approximate Message Passing for the committee machine

Input: vector $Y \in \mathbb{R}^m$ and matrix $X \in \mathbb{R}^{m\times n}$:
Initialize: $g_{\text{out},\mu} = 0, \Sigma_i = I_{K\times K}$ for $1 \le i \le n$ and $1 \le \mu \le m$ at $t = 0$.
Initialize: $\hat{W}_i \in \mathbb{R}^K$ and $\hat{C}_i, \partial_\omega g_{\text{out},\mu} \in \mathcal{S}_K^+$ for $1 \le i \le n$ and $1 \le \mu \le m$ at $t = 1$.
repeat
Update of the mean $\omega_\mu \in \mathbb{R}^K$ and covariance $V_\mu \in \mathcal{S}_K^+$:

$$\omega_\mu^t = \sum_{i=1}^n \left(\frac{X_{\mu i}}{\sqrt{n}}\hat{W}_i^t - \frac{X_{\mu i}^2}{n}\left(\Sigma_i^{t-1}\right)^{-1}\hat{C}_i^t\Sigma_i^{t-1}g_{\text{out},\mu}^{t-1}\right) \quad | \quad V_\mu^t = \sum_{i=1}^n \frac{X_{\mu i}^2}{n}\hat{C}_i^t$$

Update of $g_{\text{out},\mu} \in \mathbb{R}^K$ and $\partial_\omega g_{\text{out},\mu} \in \mathcal{S}_K^+$:

$$g_{\text{out},\mu}^t = g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t) \quad | \quad \partial_\omega g_{\text{out},\mu}^t = \partial_\omega g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t)$$

Update of the mean $T_i \in \mathbb{R}^K$ and covariance $\Sigma_i \in \mathcal{S}_K^+$:

$$T_i^t = \Sigma_i^t\Big(\sum_{\mu=1}^m \frac{X_{\mu i}}{\sqrt{n}}g_{\text{out},\mu}^t - \frac{X_{\mu i}^2}{n}\partial_\omega g_{\text{out},\mu}^t\hat{W}_i^t\Big) \quad | \quad \Sigma_i^t = -\Big(\sum_{\mu=1}^m \frac{X_{\mu i}^2}{n}\partial_\omega g_{\text{out},\mu}^t\Big)^{-1}$$

Update of the estimated marginals $\hat{W}_i \in \mathbb{R}^K$ and $\hat{C}_i \in \mathcal{S}_K^+$:

$$\hat{W}_i^{t+1} = f_w(\Sigma_i^t, T_i^t) \quad | \quad \hat{C}_i^{t+1} = f_c(\Sigma_i^t, T_i^t)$$

$t = t + 1$
until Convergence on $\hat{W}, \hat{C}$.
Output: $\hat{W}$ and $\hat{C}$.

The AMP algorithm is summarized by its pseudo-code in Algorithm 2, where the update functions g_{out}, $\partial_\omega g_{\text{out}}$, f_w and f_c are related, again, to the two auxiliary problems (3) and (4). The functions $f_w(\Sigma, T)$ and $f_c(\Sigma, T)$ are respectively the mean and variance under the posterior distribution (3) when $r \to \Sigma^{-1}$ and $Y_0 \to \Sigma^{1/2}T$, while $g_{\text{out}}(\omega_\mu, Y_\mu, V_\mu)$ is given by the product of $V_\mu^{-1/2}$ and the mean of u under the posterior (4) using $\widetilde{Y}_0 \to Y_\mu$, $\rho - q \to V_\mu$ and $q^{1/2}V \to \omega_\mu$ (see sec. F for more details). After convergence, $\hat{W}$ estimates the weights of the teacher-neural network. The label of a sample X_{new} not seen in the training set is estimated by the AMP algorithm as

$$Y_{\text{new}}^t = \int dy\Big(\prod_{l=1}^K dz_l\Big)\, y\, P_{\text{out}}(y|\{z_l\}_{l=1}^K)\mathcal{N}(z; \omega_{\text{new}}^t, V_{\text{new}}^t)\,, \tag{12}$$

where $\omega_{\text{new}}^t = \sum_{i=1}^n X_{\text{new},i}\hat{W}_i^t$ is the mean of the normally distributed variable $z \in \mathbb{R}^K$, and $V_{\text{new}}^t = \rho - q_{\text{AMP}}^t$ is the $K \times K$ covariance matrix (see below for the definition of q_{AMP}^t). We provide a demonstration code of the algorithm on GitHub [40].

AMP is particularly interesting because its performance can be tracked rigorously, again in the asymptotic limit when $n \to \infty$, via a procedure known as state evolution (a rigorous version of the cavity method in physics [14]), see [18]. State evolution tracks the value of the overlap between the hidden ground truth W^* and

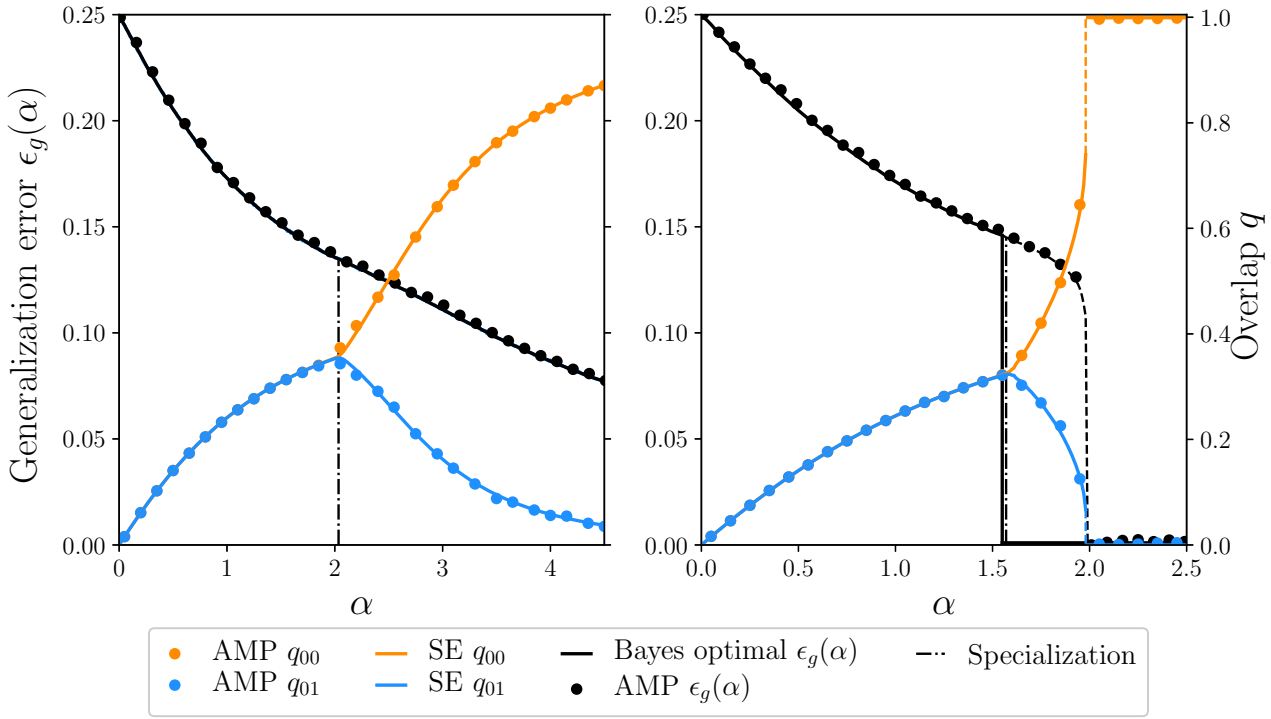

Figure 2: Generalization error and order parameter for a committee machine with two hidden neurons ($K = 2$) with Gaussian weights (left), binary/Rademacher weights (right). These are shown as a function of the ratio $\alpha = m/n$ between the number of samples m and the dimensionality n. Lines are obtained from the state evolution (SE) equations (dominating solution is shown in full line), data-points from the AMP algorithm averaged over 10 instances of the problem of size $n = 10^4$. q_{00} and q_{01} denote diagonal and off-diagonal overlaps, and their values are given by the labels on the far-right of the figure.

the AMP estimate $\hat{W}^t$, defined as $q^t_{\text{AMP}} \equiv \lim_{n\to\infty}(\hat{W}^t)^\intercal W^*/n$, via the iteration of the following equations:

$$q^{t+1}_{\text{AMP}} = 2\nabla\psi_{P_0}(r^t_{\text{AMP}})\,, \qquad r^{t+1}_{\text{AMP}} = 2\alpha\nabla\Psi_{P_{\text{out}}}(q^t_{\text{AMP}};\rho)\,. \tag{13}$$

See sec. G for more details and note that the fixed points of these equations correspond to the critical points of the replica free entropy (10).

Let us comment further on the convergence of the algorithm. In the large n limit, and if the integrals are performed without errors, then the algorithm is guaranteed to converge. This is a consequence of the state evolution combined with the Bayes-optimal setting. In practice, of course, n is finite and integrals are approximated. In that case convergence is not guaranteed, but is robustly achieved in all the cases presented in this paper. We also expect (by experience with the single layer case) that if the input-data matrix is not random (which is beyond our assumptions) then we will encounter convergence issues, which could be fixed by moving to some variant of the algorithm such as VAMP [41].

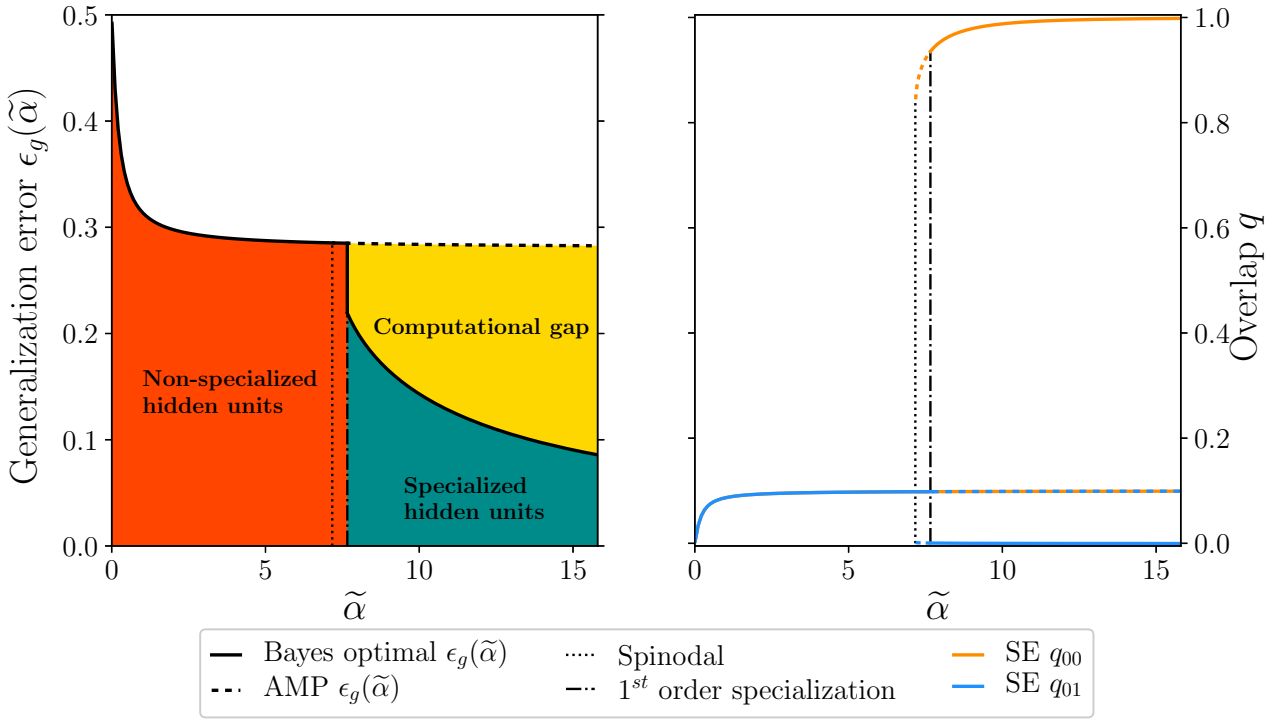

Figure 3: (Left) Bayes optimal and AMP generalization errors and (right) diagonal and off-diagonal overlaps q_{00}, q_{01} for a committee machine with a large number of hidden neurons K and Gaussian weights, as a function of the rescaled parameter $\tilde{\alpha} = \alpha/K$. Solutions corresponding to global and local minima of the replica free entropy are respectively represented with full and dashed lines. The dotted line marks the spinodal at $\widetilde{\alpha}^G_{\text{spinodal}} \simeq 7.17$, ie the apparition of a local minimum in the replica free entropy, associated to a solution with specialized hidden units. The dotted-dashed line shows the first order specialization transition at $\widetilde{\alpha}^G_{\text{spec}} \simeq 7.65$, at which the specialized fixed point becomes the global minimum. For $\widetilde{\alpha} < \widetilde{\alpha}^G_{\text{spec}}$, AMP reaches the Bayes optimal generalization error and overlaps, corresponding to a non-specialized solution. However, for $\widetilde{\alpha} > \widetilde{\alpha}^G_{\text{spec}}$, the AMP algorithm does not follow the optimal specialized solution and is stuck in the non-specialized solution plateau, represented with dashed lines. Hence it unveils a large computational gap (yellow area).

4 From two to more hidden neurons, and the specialization phase transition

4.1 Two neurons

Let us now discuss how the above results can be used to study the optimal learning in the simplest non-trivial case of a two-layers neural network with two hidden neurons, that is when model (1) is simply

$$Y_\mu = \text{sign}\Big[\,\text{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W^*_{i1}\Big) + \text{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W^*_{i2}\Big)\Big]\,,$$

and is represented in Fig. 1, with the convention that $\text{sign}(0) = 0$. We remind that the input-data matrix X has i.i.d. $\mathcal{N}(0,1)$ entries, and the teacher-weights W^* used to generate the labels Y are taken i.i.d. from P_0.

In Fig. 2 we plot the optimal generalization error as a function of the sample complexity $\alpha = m/n$. In the left panel the weights are Gaussian (for both the teacher and the student), while in the right panel they are binary/Rademacher. The full line is obtained from the fixed point of the state evolution (SE) of the AMP algorithm (13), corresponding to the extremizer of the replica free entropy (10). The points are results of the AMP algorithm run till convergence averaged over 10 instances of size $n = 10^4$. In this case and with random initial conditions the AMP algorithm did converge in all our trials. As expected we observe excellent

agreement between the SE and AMP.

In both left and right panels of Fig. 2 we observe the so-called *specialization* phase transition. Indeed (13) has two types of fixed points: a *non-specialized* fixed point where every matrix element of the $K \times K$ order parameter q is the same (so that both hidden neurons learn the same function) and a *specialized* fixed point where the diagonal elements of the order parameter are different from the non-diagonal ones. We checked for other types of fixed points for $K = 2$ (one where the two diagonal elements are not the same), but have not found any. In terms of weight-learning, this means for the non-specialized fixed point that the estimators for both W_1 and W_2 are the same, whereas in the specialized fixed point the estimators of the weights corresponding to the two hidden neurons are different, and that the network "figured out" that the data are better described by a model that is not linearly separable. The specialized fixed point is associated with lower error than the non-specialized one (as one can see in Fig. 2). The existence of this phase transition was discussed in statistical physics literature on the committee machine, see e.g. [20, 23].

For Gaussian weights (Fig. 2 left), the specialization phase transition arises continuously at $\alpha^G_{\rm spec}(K = 2) \simeq 2.04$. This means that for $\alpha < \alpha^G_{\rm spec}(K = 2)$ the number of samples is too small, and the student-neural network is not able to learn that two different teacher-vectors W_1 and W_2 were used to generate the observed labels. For $\alpha > \alpha^G_{\rm spec}(K = 2)$, however, it is able to distinguish the two different weight-vectors and the generalization error decreases fast to low values (see Fig. 2). For completeness we remind that in the case of $K = 1$ corresponding to single-layer neural network no such specialization transition exists. We show in sec. E that it is absent also in multi-layer neural networks as long as the activations remain linear. The non-linearity of the activation function is therefore an essential ingredient in order to observe a specialization phase transition.

The right part of Fig. 2 depicts the fixed point reached by the state evolution of AMP for the case of binary weights. We observe two phase transitions in the performance of AMP in this case: (a) the specialization phase transition at $\alpha^B_{\rm spec}(K = 2) \simeq 1.58$, and for slightly larger sample complexity a transition towards *perfect generalization* (beyond which the generalization error is asymptotically zero) at $\alpha^B_{\rm perf}(K = 2) \simeq 1.99$. The binary case with $K = 2$ differs from the Gaussian one in the fact that perfect generalization is achievable at finite α. While the specialization transition is continuous here, the error has a discontinuity at the transition of perfect generalization. This discontinuity is associated with the 1st order phase transition (in the physics nomenclature), leading to a gap between algorithmic (AMP in our case) performance and information-theoretically optimal performance reachable by exponential algorithms. To quantify the optimal performance we need to evaluate the global extremum of the replica free entropy (not the local one reached by the state evolution). In doing so that we get that information theoretically there is a single discontinuous phase transition towards perfect generalization at $\alpha^B_{\rm IT}(K = 2) \simeq 1.54$.

While the information-theoretic and specialization phase transitions were identified in the physics literature on the committee machine [20, 21, 3, 4], the gap between the information-theoretic performance and the performance of AMP —that is conjectured to be optimal among polynomial algorithms— was not yet discussed in the context of this model. Indeed, even its understanding in simpler models than those discussed here, such as the single layer case, is more recent [15, 26, 25].

4.2 More is different

It becomes more difficult to study the replica formula for larger values of K as it involves (at least) K-dimensional integrals. Quite interestingly, it is possible to work out the solution of the replica formula in the large K limit (thus taken *after* the large n limit, so that K/n vanishes). It is indeed natural to look for solutions of the replica formula, as suggested in [19], of the form $q = q_d I_{K\times K} + (q_a/K)\mathbf{1}_K\mathbf{1}_K^\intercal$, with the unit vector $\mathbf{1}_K = (1)_{l=1}^K$. Since both q and ρ are assumed to be positive, this scaling implies that $0 \leq q_d \leq 1$ and $0 \leq q_a + q_d \leq 1$, as it should, see sec. D. We also detail in this same section the corresponding large K

expansion of the free entropy for the teacher-student scenario with Gaussian weights. Only the information-theoretically reachable generalization error was computed [19], thus we concentrated on the analysis of performance of AMP by tracking the state evolution equations. In doing so, we unveil a large computational gap.

In the right panel of Fig. 3 we show the fixed point values of the two overlaps $q_{00} = q_d + q_a/K$ and $q_{01} = q_a/K$ and the resulting generalization error, plotted in the left panel. As discussed in [19] it can be written in a closed form as $\epsilon_g = \arccos\left[2\left(q_a + \arcsin q_d\right)/\pi\right]/\pi$, represented in the left panel of Fig. 3. The specialization transition arises for $\alpha = \Theta(K)$ so we define $\widetilde{\alpha} \equiv \alpha/K$. The specialization is now a 1st order phase transition, meaning that the specialization fixed point first appears at $\widetilde{\alpha}^G_{\text{spinodal}} \simeq 7.17$ but the free entropy global extremizer remains the one of the non-specialized fixed point until $\widetilde{\alpha}^G_{\text{spec}} \simeq 7.65$. This has interesting implications for the optimal generalization error that gets towards a plateau of value $\varepsilon_{\text{plateau}} \simeq 0.28$ for $\widetilde{\alpha} < \widetilde{\alpha}^G_{\text{spec}}$ and then jumps discontinuously down to reach a decay aymptotically as $1.25/\widetilde{\alpha}$. See left panel of Fig. 3.

AMP is conjectured to be optimal among all polynomial algorithms (in the considered limit) and thus analyzing its state evolution sheds light on possible computational-to-statistical gaps that come hand in hand with 1st order phase transitions. In the regime of $\alpha = \Theta(K)$ for large K the non-specialized fixed point is always stable implying that AMP will not be able to give a lower generalization error than $\varepsilon_{\text{plateau}}$. Analyzing the replica formula for large K in more details, see sec. D, we concluded that AMP will not reach the optimal generalization for any $\alpha < \Theta(K^2)$. This implies a rather sizable gap between the performance that can be reached information-theoretically and the one reachable tractably (see yellow area in Fig. 3). Such large computational gaps have been previously identified in a range of inference problems —most famously in the planted clique problem [27]— but the committee machine is the first model of a multi-layer neural network with realistic non-linearities (the parity machine is another example but use a very peculiar non-linearity) that presents such large gap.

5 Structure of the proof of Theorem 3.1

All along this section we assume (H1), (H2) and (H3), and all the rigorous statements are implicitely assuming these hypotheses. We denote K-dimensional column vectors by underlined letters. In particular $\underline{W}^*_i = (W^*_{il})_{l=1}^K, \underline{w}_i = (w_{il})_{l=1}^K$. For $\mu = 1, \dots m$, let $\underline{V}_\mu, \underline{U}^*_\mu$ be K-dimensional vectors with i.i.d. $\mathcal{N}(0,1)$ components. Let $s_n \in (0, 1/2]$ a sequence that goes to 0 as n increases, and let $\mathcal{M}$ be the compact subset of matrices in S_K^{++} with eigenvalues in the interval $[1, 2]$. For all $M \in s_n\mathcal{M}$, $2s_n I_{K\times K} - M \in \mathcal{S}_K^+$.

5.1 Interpolating estimation problem

Let $\epsilon = (\epsilon_1, \epsilon_2) \in (s_n\mathcal{M})^2$. Let $q : [0,1] \to \mathcal{S}_K^+(\rho)$ and $r : [0,1] \to \mathcal{S}_K^+$ be two "interpolation functions" (that will later on depend on ϵ), and

$$R_1(t) \equiv \epsilon_1 + \int_0^t r(v)dv\,, \qquad R_2(t) \equiv \epsilon_2 + \int_0^t q(v)dv\,. \tag{14}$$

For $t \in [0,1]$, define the K-dimensional vector:

$$\underline{S}_{t,\mu} \equiv \sqrt{\frac{1-t}{n}} \sum_{i=1}^n X_{\mu i}\underline{W}^*_i + \sqrt{R_2(t)}\,\underline{V}_\mu + \sqrt{t\rho - R_2(t) + 2s_n I_{K\times K}}\,\underline{U}^*_\mu \tag{15}$$

where matrix square-roots (that we denote equivalently $A^{1/2}$ or $\sqrt{A}$) are well defined. We interpolate with auxiliary problems related to those discussed in sec. 3; the interpolating estimation problem is given by the

following observation model, with two types of t-dependent observations:

$$\begin{cases} Y_{t,\mu} \sim P_{\rm out}(\,\cdot\,|\,\underline{S}_{t,\mu}), & 1 \le \mu \le m\,, \\ \underline{Y}'_{t,i} = \sqrt{R_1(t)}\,\underline{W}^*_i + \underline{Z}'_i, & 1 \le i \le n\,, \end{cases} \tag{16}$$

where $\underline{Z}'_i$ is (for each i) a K-vector with i.i.d. $\mathcal{N}(0,1)$ components, and $\underline{Y}'_{t,i}$ is a K-vector as well. Recall that in our notation the $*$-variables have to be retrieved, while the other random variables are assumed to be known (except for the noise variables obviously). Define now $\underline{s}_{t,\mu}$ by the expression of $\underline{S}_{t,\mu}$ but with $\underline{w}_i$ replacing $\underline{W}^*_i$ and $\underline{u}_\mu$ replacing $\underline{U}^*_\mu$. We introduce the *interpolating posterior*:

$$P_{t,\epsilon}(w,u|Y_t,Y'_t,X,V) = \frac{1}{\mathcal{Z}_{n,\epsilon}(t)} \prod_{i=1}^n P_0(\underline{w}_i) e^{-\frac{1}{2}\|\underline{Y}'_{t,i} - \sqrt{R_1(t)}\underline{w}_i\|_2^2} \prod_{\mu=1}^m \frac{e^{-\frac{1}{2}\|\underline{u}_\mu\|_2^2}}{(2\pi)^{K/2}} P_{\rm out}(Y_{t,\mu}|\underline{s}_{t,\mu}) \tag{17}$$

where the normalization factor $\mathcal{Z}_{n,\epsilon}(t)$ equals the numerator integrated over all components of w and u. The average free entropy at time t is by definition

$$f_{n,\epsilon}(t) \equiv \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_{n,\epsilon}(t) = \frac{1}{n}\mathbb{E}\ln \int \mathcal{D}u \prod_{i=1}^n dP_0(\underline{w}_i) \prod_{\mu=1}^m P_{\rm out}(Y_{t,\mu}|\underline{s}_{t,\mu}) \prod_{i=1}^n e^{-\frac{1}{2}\|\underline{Y}'_{t,i} - \sqrt{R_1(t)}\underline{w}_i\|_2^2}\,, \tag{18}$$

where $\mathcal{D}u = \prod_{\mu=1}^m \prod_{l=1}^K (2\pi)^{-1/2} e^{-u_{\mu l}^2/2}$.

The presence of the small "pertubation" ϵ induces a proportional change in the free entropy of the interpolating model:

Lemma 5.1 (Perturbation of the free entropy). *For all $\epsilon \in (s_n\mathcal{M})^2$ we have for $t = 0$ that $|f_{n,\epsilon}(0) - f_{n,\epsilon=(0,0)}(0)| \le C' s_n$ for some positive constant C'. Moreover, $|f_n - f_{n,\epsilon=(0,0)}(0)| \le C s_n$ for some positive constant C, so that*

$$|f_n - f_{n,\epsilon=(0,0)}(0)| = \mathcal{O}_n(1)\,.$$

Proof. Let us compute (or directly obtain by the I-MMSE formula for vector channels [42, 43, 44])

$$\nabla_{\epsilon_1} f_{n,\epsilon}(0) = -\frac{1}{2}\left[\rho - \mathbb{E}\langle Q\rangle_{n,0,\epsilon}\right], \tag{19}$$

where the $K \times K$ *overlap matrix* $(Q_{ll'})$ is defined below by (23). Note that the r.h.s. of the above equation is (up to a factor $-1/2$) the $K \times K$ MMSE matrix. Set $u_y(x) \equiv \ln P_{\rm out}(y|x)$. Now we compute (by calculations very similar to the ones used in the proof of the following Proposition 5.2):

$$\nabla_{\epsilon_2} f_{n,\epsilon}(0) = \frac{1}{2n}\sum_{\mu=1}^m \mathbb{E}\Big[\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\Big\langle \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\Big\rangle_{n,0,\epsilon}\Big]. \tag{20}$$

Note that the r.h.s. of the above equation is symmetric by the Nishimori identity Proposition A.1. By the mean value theorem we obtain then directly that $|f_{n,\epsilon}(0) - f_{n,\epsilon=(0,0)}(0)| \le \|\nabla_{\epsilon_1} f_{n,\epsilon}(0)\|_{\rm F}\|\epsilon_1\|_{\rm F} + \|\nabla_{\epsilon_2} f_{n,\epsilon}(0)\|_{\rm F}\|\epsilon_2\|_{\rm F} \le C \max_i \|\epsilon_i\| \le C' s_n$. □

Using this lemma one verifies, using in particular continuity and boundedness properties of ψ_{P_0} and $\Psi_{P_{\rm out}}$ (see Lemma A.6 in sec. A for details; sec. A gathers the detailed proofs of all the propositions below):

$$\begin{cases} f_{n,\epsilon}(0) &= f_n - \frac{K}{2} + \mathcal{O}_n(1)\,, \\ f_{n,\epsilon}(1) &= \psi_{P_0}(\int_0^1 r(t)dt) + \alpha\Psi_{P_{\rm out}}(\int_0^1 q(t)dt;\rho) - \frac{1}{2}\int_0^1 {\rm Tr}[\rho\, r(t)]dt - \frac{K}{2} + \mathcal{O}_n(1)\,. \end{cases} \tag{21}$$

Here $\mathcal{O}_n(1) \to 0$ in the $n, m \to \infty$ limit uniformly in t, q, r, ϵ.

5.2 Overlap concentration and fundamental sum rule

Notice from (21) that at $t = 1$ the interpolating estimation problem constructs part of the RS potential (9), while at $t = 0$ it is the free entropy (6) of the original model (7) (up to a constant). We thus now want to compare these boundary values thanks to the identity

$$f_n = f_{n,\epsilon}(0) + \frac{K}{2} + \mathcal{O}_n(1) = f_{n,\epsilon}(1) - \int_0^1 \frac{df_{n,\epsilon}(t)}{dt} dt + \frac{K}{2} + \mathcal{O}_n(1)\,. \tag{22}$$

The next obvious step is therefore to compute the free entropy variation along the interpolation path, see sec. A.3 for the proof:

Proposition 5.2 (Free entropy variation). *Denote by $\langle - \rangle_{n,t,\epsilon}$ the (Gibbs) expectation w.r.t. the posterior $P_{t,\epsilon}$ given by (17). Set $u_y(x) \equiv \ln P_{\rm out}(y|x)$. For all $t \in [0,1]$ we have*

$$\frac{df_{n,\epsilon}(t)}{dt} = -\frac{1}{2}\mathbb{E}\Big\langle \mathrm{Tr}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - r(t)\Big)\big(Q - q(t)\big)\Big]\Big\rangle_{n,t,\epsilon} + \frac{1}{2}\mathrm{Tr}\left[r(t)(q(t) - \rho)\right] + \mathcal{O}_n(1)\,,$$

where ∇ is the K-dimensional gradient w.r.t. the argument of $u_{Y_{t,\mu}}(\cdot)$, and $\mathcal{O}_n(1) \to 0$ in the $n, m \to \infty$ limit uniformly in t, q, r, ϵ. Here, the $K \times K$ overlap matrix Q is defined as

$$Q_{ll'} \equiv \frac{1}{n}\sum_{i=1}^{n} W^*_{il} w_{il'}\,. \tag{23}$$

We will plug this expression in identity (22), but in order to simplify it we need the following crucial proposition, which says that the overlap concentrates. This property is what is generally refered to as a replica symmetric behavior in statistical physics.

Proposition 5.3 (Overlap concentration). *Assume that for any $t \in (0,1)$ the transformation $\epsilon \in (s_n\mathcal{M})^2 \mapsto (R_1(t,\epsilon), R_2(t,\epsilon))$ is a $\mathcal{C}^1$ diffeomorphism with a Jacobian determinant greater or equal to 1. Then one can find a sequence s_n going to 0 slowly enough such that there exists a constant $C(\varphi_{\rm out}, S, K, \alpha) > 0$ depending only on the activation $\varphi_{\rm out}$, the support S of the prior P_0, the number of hidden neurons K and the sampling rate α, and a constant $\gamma > 0$ such that ($\|-\|_{\rm F}$ is the Frobenius norm):*

$$\frac{1}{\mathrm{Vol}(s_n\mathcal{M})^2}\int_{(s_n\mathcal{M})^2} d\epsilon \int_0^1 dt\, \mathbb{E}\big\langle \big\|Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon}\big\|_{\rm F}^2\big\rangle_{n,t,\epsilon} \leq \frac{C(\varphi_{\rm out}, S, K, \alpha)}{n^\gamma}\,.$$

The proof of this concentration result can be directly adapted from [45]. Using the results of [45] is straightforward, under the assumption that $\epsilon \mapsto R(t,\epsilon)$ is a $\mathcal{C}^1$ diffeomorphism with a Jacobian determinant greater or equal to 1. This Jacobian determinant can be computed from formula (30). To check that it is greater than one we use Lemma 5.5 and need Assumption 1 stated in paragraph 5.3 below. With a Jacobian determinant greater than one, we can "replace" (i.e., lower bound) the integrations over $R_1(t,\epsilon)$, that naturally appear in the proof of Proposition 5.3, by integrations over the perturbation matrix ϵ. This is *exactly* what has been done in the $K = 1$ version of the present model in [11] or in [46] i.e., in the scalar overlap case (see also [47] for a setting with a matrix overlap as in the present case).

From there we can deduce the following fundamental sum rule which is at the core of the proof:

Proposition 5.4 (Fundamental sum rule). *Assume that the interpolation functions r and q are such that the map $\epsilon = (\epsilon_1, \epsilon_2) \mapsto R(t,\epsilon) = (R_1(t,\epsilon), R_2(t,\epsilon))$ given by (14) is a $\mathcal{C}^1$ diffeomorphism whose Jacobian determinant*

$J_{n,\epsilon}(t)$ is greater or equal to 1. Assume that for all $t \in [0,1]$ and $\epsilon \in (s_n\mathcal{M})^2$ we have $q(t) = q(t,\epsilon) = \mathbb{E}\langle Q\rangle_{n,t,\epsilon} \in \mathcal{S}_K^+(\rho)$. Then

$$f_n = \frac{1}{\mathrm{Vol}(s_n\mathcal{M})^2}\int_{(s_n\mathcal{M})^2} d\epsilon \Big\{ \psi_{P_0}\Big(\int_0^1 r(t)dt\Big) + \alpha\Psi_{P_{\rm out}}\Big(\int_0^1 q(t,\epsilon)dt;\rho\Big) - \frac{1}{2}\int_0^1 \mathrm{Tr}[q(t,\epsilon)r(t)]dt\Big\} + \mathcal{O}_n(1)\,. \tag{24}$$

Proof. Let us denote $V_n \equiv \mathrm{Vol}(s_n\mathcal{M})^2$. The integral over ϵ is always over $(s_n\mathcal{M})^2$. Consider the first term, i.e. the Gibbs bracket, in the free entropy derivative given by Proposition 5.2. By the Cauchy-Schwarz inequality

$$\Big(\mathbb{E}\Big\langle \mathrm{Tr}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^m \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - r(t)\Big)\big(Q-q(t)\big)\Big]\Big\rangle_{n,t,\epsilon}\Big)^2$$
$$\leq \frac{1}{V_n}\int d\epsilon\int_0^1 dt\,\mathbb{E}\Big\langle\Big\|\frac{1}{n}\sum_{\mu=1}^m \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - r(t)\Big\|_{\rm F}^2\Big\rangle_{n,t,\epsilon} \times \frac{1}{V_n}\int d\epsilon\int_0^1 dt\,\mathbb{E}\big\langle\big\|Q-q(t)\big\|_{\rm F}^2\big\rangle_{n,t,\epsilon}\,.$$

The first term of this product is bounded by some constant $C(\varphi_{\rm out},\alpha)$ that only depend on $\varphi_{\rm out}$ and α, see Lemma A.4 in sec. A.4. The second term is bounded by $C(\varphi_{\rm out},S,K,\alpha)n^{-\gamma}$ by Proposition 5.3, since we assumed that for all $\epsilon \in \mathcal{B}_n$ and all $t \in [0,1]$ we have $q(t) = q(t,\epsilon) = \mathbb{E}\langle Q\rangle_{n,t,\epsilon}$. Therefore from Proposition 5.2 we obtain

$$\frac{1}{V_n}\int d\epsilon\int_0^1 \frac{df_{n,\epsilon}(t)}{dt}dt = \frac{1}{2V_n}\int d\epsilon\int_0^1 \mathrm{Tr}\big[q(t,\epsilon)r(t) - r(t)\rho\big]dt + \mathcal{O}_n(1) + \mathcal{O}(n^{-\gamma/2})\,. \tag{25}$$

Here the small terms are both going to 0 uniformly w.r.t. to the choice of q and r. When replacing (25) in (22) and combining it with (21) we reach the claimed identity. □

5.3 A technical lemma and an assumption

We give here a technical lemma used in the rest of the proof, and which allows us to detail the unproven assumption on which we rely to prove Thm 3.1.

Lemma 5.5. *The quantity $\mathbb{E}\langle Q\rangle_{n,t,\epsilon}$ is a function of $(n,t,R(t,\epsilon))$. We define $F_n^{(1)}(t,R(t,\epsilon)) \equiv \mathbb{E}\langle Q\rangle_{n,t,\epsilon}$ and $F_n^{(2)}(t,R(t,\epsilon)) \equiv 2\alpha\nabla\Psi_{P_{\rm out}}(\mathbb{E}\langle Q\rangle_{n,t,\epsilon})$. $F_n \equiv (F_n^{(1)},F_n^{(2)})$ is defined on the set:*

$$D_n = \Big\{(t,r_1,r_2) \in [0,1]\times\mathcal{S}_K^+\times\mathcal{S}_K^+\Big|(\rho t - r_2 + 2s_nI_K) \in \mathcal{S}_K^+\Big\}. \tag{26}$$

F_n is a continuous function from D_n to $\mathcal{S}_K^+\times\mathcal{S}_K^+(\rho)$. Moreover, F_n admits partial derivatives with respect to R_1 and R_2 on the interior of D_n. For every $(t,R(t,\epsilon))$ for which they are defined, they satisfy:

$$\sum_{l\leq l'}^K \frac{\partial(F_n^{(1)})_{ll'}}{\partial(R_1)_{ll'}} \geq 0. \tag{27}$$

We can now state the technical assumption on which we rely, and which essentially allows us to derive that the map $\epsilon \mapsto R(t,\epsilon)$ is a $\mathcal{C}^1$ diffeomorphism with a Jacobian determinant greater or equal to 1 as it will become clear in the next section:

Assumption 1. *With the notations of Lemma 5.5,*

$$\sum_{l\le l'}^{K} \frac{\partial (F_n^{(2)})_{ll'}}{\partial (R_2)_{ll'}} \ge 0.$$

Proof of Lemma 5.5. The fact that the image domain of F_n is $\mathcal{S}_K^+ \times \mathcal{S}_K^+(\rho)$ is known from Lemma A.2. The continuity and differentiability of F_n follows from standard theorems of continuity and derivation under the integral sign (recall that we are working at finite n). Indeed, the domination hypotheses are easily satisfied since we work under (H1) and (H2).

Let us now prove (27). We write the formal differential of $F_n^{(1)}$ with respect to R_1 as $\mathcal{D}_{R_1} F_n^{(1)}$, which is a 4-tensor, and our goal is to prove that $\mathrm{Tr}[\mathcal{D}_{R_1} F_n^{(1)}] \ge 0$, the trace of a 4-tensor over S_K $A_{(ij)(kl)}$ being $\mathrm{Tr}[A] = \sum_{i\le j} A_{(ij)(ij)}$. Then one can write $\mathrm{Tr}[\mathcal{D}_{R_1} F_n^{(1)}] = \mathrm{Tr}[\nabla\nabla^\intercal \Psi_{P_{\mathrm{out}}}(\mathbb{E}\langle Q\rangle_{n,t,\epsilon}) \times \nabla_{R_1} \mathbb{E}\langle Q\rangle_{n,t,\epsilon}]$. We know from Lemma A.2 and Lemma A.6 that $\nabla\nabla^\intercal \Psi_{P_{\mathrm{out}}}(\mathbb{E}\langle Q\rangle_{n,t,\epsilon})$ is a positive symmetric matrix (when seen as a linear operator over $\mathcal{S}_K$). Moreover, it is a known result that the derivative $\nabla_{R_1} \mathbb{E}\langle Q\rangle_{n,t,\epsilon}$ is also positive symmetric, since R_1 is the matrix snr of a *linear* channel (see [42, 43, 44]). Since the product of two symmetric positive matrices has always positive trace, this shows that $\mathrm{Tr}[\mathcal{D}_{R_1} F_n^{(1)}] \ge 0$. □

5.4 Matching bounds

Proposition 5.6 (Lower bound). *Under Assumption 1, the free entropy of model (7) verifies*

$$\liminf_{n\to\infty} f_n \ge \sup_{r\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q,r)\,.$$

Proof. Choose first $r(t) = r \in \mathcal{S}_K^+$ a fixed matrix. Then $R(t) = (R_1(t), R_2(t))$ can be fixed as the solution to the first order differential equation:

$$\frac{d}{dt} R_1(t) = r\,, \qquad \frac{d}{dt} R_2(t) = \mathbb{E}\langle Q\rangle_{n,t,\epsilon}\,, \qquad \text{and} \qquad R(0) = \epsilon\,. \tag{28}$$

We denote this (unique) solution $R(t,\epsilon) = (rt + \epsilon_1, \int_0^t q(v,\epsilon;r)dv + \epsilon_2)$. It is possible to check that this ODE satisfies the hypotheses of the parametric Cauchy-Lipschitz theorem, and that by the Liouville formula the determinant $J_{n,\epsilon}(t)$ of the Jacobian of $\epsilon \mapsto R(t,\epsilon)$ satisfies (see Lemma A.3 in sec. A)

$$J_{n,\epsilon}(t) = \exp\Big(\int_0^t \sum_{l\ge l'}^{K} \frac{\partial \mathbb{E}\langle Q_{ll'}\rangle_{n,s,\epsilon}}{\partial (R_2)_{ll'}}(s, R(s,\epsilon))\, ds \Big) \ge 1\,. \tag{29}$$

Indeed, this sum of partial derivatives is always positive by Assumption 1. Moreover from (28), $q(t,\epsilon;r) = \mathbb{E}\langle Q\rangle_{n,t,\epsilon}$, which is in $\mathcal{S}_K^+$ by Lemma A.2 in sec. A. The fact that the map $\epsilon \mapsto R(t,\epsilon)$ is a $\mathcal{C}^1$ diffeomorphism is easily verified by its bijectivity (from the positivity of $J_{n,\epsilon}(t)$) combined with the local inversion Theorem. All the assumptions of Proposition 5.4 are veri.i.d. which then implies, recalling the potential expression (9),

$$f_n = \frac{1}{\mathrm{Vol}(s_n\mathcal{M})^2} \int_{(s_n\mathcal{M})^2} d\epsilon\, f_{\mathrm{RS}}\Big(\int_0^1 q(v,\epsilon;r)dv, r \Big) + \mathcal{O}_n(1)\,.$$

This implies the lower bound as this equality is true for any $r \in \mathcal{S}_K^+$. □

Proposition 5.7 (Upper bound). *Under Assumption 1, the free entropy of model (7) verifies*

$$\limsup_{n\to\infty} f_n \le \sup_{r\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\rm RS}(q,r)\,.$$

Proof. We now fix $R(t)=(R_1(t),R_2(t))$ as the solution $R(t,\epsilon)=(\int_0^t r(v,\epsilon)dv+\epsilon_1,\int_0^t q(v,\epsilon)dv+\epsilon_2)$ to the following Cauchy problem:

$$\frac{d}{dt}R_1(t)=2\alpha\nabla\Psi_{P_{\rm out}}(\mathbb{E}\langle Q\rangle_{n,t,\epsilon})\,,\qquad \frac{d}{dt}R_2(t)=\mathbb{E}\langle Q\rangle_{n,t,\epsilon}\,,\qquad\text{and}\qquad R(0)=\epsilon\,.$$

We denote this equation as $\partial_t R(t)=F_n(t,R(t)),R(0)=\epsilon$. It is then possible to verify that $F_n(R(t),t)$ is a bounded $\mathcal{C}^1$ function of $R(t)$, and thus a direct application of the Cauchy-Lipschitz theorem implies that $R(t,\epsilon)$ is a $\mathcal{C}^1$ function of t and ϵ. The Liouville formula for the Jacobian determinant of the map $\epsilon\in(s_n\mathcal{M})^2\mapsto R(t,\epsilon)\in R(t,(s_n\mathcal{M})^2)$ gives this time (see Lemma A.3 in sec. A)

$$J_{n,\epsilon}(t)=\exp\Big(\int_0^t\sum_{l\ge l'}^{K}\Big\{\frac{\partial(F_{n,1})_{ll'}}{\partial(R_1)_{ll'}}(s,R(s,\epsilon))+\frac{\partial(F_{n,2})_{ll'}}{\partial(R_2)_{ll'}}(s,R(s,\epsilon))\Big\}ds\Big)\ge 1\,. \tag{30}$$

The fact that this determinant is greater or equal to 1 for all $t\in[0,1]$ follows again from the positivity of this sum of partials, see Lemma 5.5 and Assumption 1. Identity (30) implies the bijectivity of $\epsilon\mapsto R(t,\epsilon)$ which, combined with the local inversion theorem, makes it a diffeomorphism. Since $\mathbb{E}\langle Q\rangle_{n,t,\epsilon}$ and $\rho-\mathbb{E}\langle Q\rangle_{n,t,\epsilon}$ are positive matrices (see Lemma A.2 in sec. A) we also have that $q(t,\epsilon)\in\mathcal{S}_K^+(\rho)$ and since by the differential equation we have $r(t,\epsilon)=2\alpha\nabla\Psi_{P_{\rm out}}(q(t,\epsilon))$ and as $\nabla\Psi_{P_{\rm out}}(q)\in\mathcal{S}_K^+$ (see Lemma A.6 in sec. A), then $r(t,\epsilon)\in\mathcal{S}_K^+$ too. We have everything needed for applying Proposition 5.4 again which gives in this case

$$f_n=\frac{1}{{\rm Vol}(s_n\mathcal{M})^2}\int d\epsilon\Big\{\psi_{P_0}\Big(\int_0^1 r(v,\epsilon)dv\Big)+\alpha\Psi_{P_{\rm out}}\Big(\int_0^1 q(v,\epsilon)dv;\rho\Big)-\frac{1}{2}{\rm Tr}\int_0^1 q(v,\epsilon)r(v,\epsilon)dv\Big\}+\mathcal{O}_n(1).$$

Then by convexity of ψ_{P_0} and $\Psi_{P_{\rm out}}$ (see Lemma A.6),

$$\begin{aligned}f_n&\le\frac{1}{{\rm Vol}(s_n\mathcal{M})^2}\int d\epsilon\int_0^1 dv\Big\{\psi_{P_0}(r(v,\epsilon)dv)+\alpha\Psi_{P_{\rm out}}(q(v,\epsilon);\rho)-\frac{1}{2}{\rm Tr}[q(v,\epsilon)r(v,\epsilon)]\Big\}+\mathcal{O}_n(1)\\&=\frac{1}{{\rm Vol}(s_n\mathcal{M})^2}\int d\epsilon\int_0^1 dv\,f_{\rm RS}(q(v,\epsilon),r(v,\epsilon))+\mathcal{O}_n(1)\,.\end{aligned}$$

We now remark that

$$f_{\rm RS}(q(v,\epsilon),r(v,\epsilon))=\inf_{q\in\mathcal{S}_K^+(\rho)}f_{\rm RS}(q,r(v,\epsilon))\,.$$

Indeed, for every $r\in\mathcal{S}_K^+$, the function $g_r:q\in\mathcal{S}_K^+(\rho)\mapsto f_{\rm RS}(q,r)\in\mathbb{R}$ (recall (9)) is convex (by Lemma A.6), and its q-derivative is $\nabla g_r(q)=\alpha\nabla\Psi_{P_{\rm out}}(q)-r/2$. Since $\nabla g_{r(v,\epsilon)}(q(v,\epsilon))=0$ by definition of $r(v,\epsilon)$, and $\mathcal{S}_K^+(\rho)$ is convex, the minimum of $g_{r(v,\epsilon)}(q)$ is necessarily achieved at $q=q(v,\epsilon)$. Therefore

$$f_n\le\frac{1}{{\rm Vol}(s_n\mathcal{M})^2}\int_{(s_n\mathcal{M})^2}d\epsilon\int_0^1 dv\inf_{q\in\mathcal{S}_K^+(\rho)}f_{\rm RS}\,(q,r(v,\epsilon))+\mathcal{O}_n(1)\le\sup_{r\in\mathcal{S}_K^+}\inf_{q\in\mathcal{S}_K^+(\rho)}f_{\rm RS}(q,r)+\mathcal{O}_n(1),$$

which concludes the proof of Proposition 5.7. □

Combining these two matching bounds ends the proof of Theorem 3.1.

6 Discussion

One of the contributions of this paper is the design of an AMP-type algorithm that is able to achieve the Bayes-optimal learning error in the limit of large dimensions for a range of parameters out of the so-called hard phase. The hard phase is associated with first order phase transitions appearing in the solution of the model. In the case of the committee machine with a large number of hidden neurons we identify a large hard phase in which learning is possible information-theoretically but not efficiently. In other problems where such a hard phase was identified, its study boosted the development of algorithms that are able to match the predicted threshold. We anticipate this will also be the same for the present model. We should, however, note that for larger $K > 2$ the present AMP algorithm includes higher-dimensional integrals that hamper the speed of the algorithm. Our current strategy to tackle this is to combine the large-K expansion and use it in the algorithm. Detailed account of the corresponding results are left for future work.

We studied the Bayes-optimal setting where the student-network is the same as the teacher-network, for which the replica method can be readily applied. The method still applies when the number of hidden units in the student and teacher are different, while our proof does not generalize easily to this case. It is an interesting subject for future work to see how the hard phase evolves under over-parametrization and what is the interplay between the simplicity of the loss-landscape and the achievable generalization error. We conjecture that in the present model over-parametrization will not improve the generalization error achieved by AMP in the Bayes-optimal case.

Even though we focused in this paper on a two-layers neural network, the analysis and algorithm can be readily extended to a multi-layer setting, see [22], as long as the number of layers as well as the number of hidden neurons in each layer is held constant, and as long as one learns only weights of the first layer, for which the proof already applies. The numerical evaluation of the phase diagram would be more challenging than the cases presented in this paper as multiple integrals would appear in the corresponding formulas. In future works, we also plan to analyze the case where the weights of the second and subsequent layers (including the biases of the activation functions) are also learned. This could be done for instance with a combination of EM and AMP along the lines of [48, 49] where this is done for the simpler single layer case.

Concerning extensions of the present work, an important open case is the one where the number of samples per dimension $\alpha = \Theta(1)$ and also the size of the hidden layer per dimension $K/n = \Theta(1)$ as $n \to \infty$, while in this paper we treated the case $K = \Theta(1)$ and $n \to \infty$. This other scaling where $K/n = \Theta(1)$ is challenging even for the non-rigorous replica method.

Acknowledgments

This work has been supported by the ERC under the European Union's FP7 Grant Agreement 307087-SPARCS and the European Union's Horizon 2020 Research and Innovation Program 714608-SMiLe, as well as by the French Agence Nationale de la Recherche under grant ANR-17-CE23-0023-01 PAIL and the Swiss National Foundation grant no 200021E-175541. Additional funding is acknowledged by A.M., F.K. and J.B. from "Chaire de recherche sur les modèles et sciences des données", Fondation CFM pour la Recherche-ENS. We also acknowledge Léo Miolane for discussions.

References

[1] V. Vapnik. *Statistical learning theory. 1998.* Wiley, New York, 1998.

[2] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[3] S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056, 1992.

[4] T. L. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning a rule. *Reviews of Modern Physics*, 65(2):499, 1993.

[5] R. Monasson and R. Zecchina. Learning and generalization theories of large committee-machines. *Modern Physics Letters B*, 9(30):1887–1897, 1995.

[6] R. Monasson and R. Zecchina. Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks. *Physical review letters*, 75(12):2432, 1995.

[7] A. Engel and C. P. Van den Broeck. *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[8] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016. in ICLR 2017.

[9] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016. in ICLR 2017.

[10] C. H. Martin and M. W. Mahoney. Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior. *arXiv preprint arXiv:1710.09553*, 2017.

[11] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

[12] M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. Ben-Arous, C. Cammarota, Y. LeCun, M. Wyart, and G. Biroli. Comparing dynamics: Deep neural networks versus glassy systems. *arXiv preprint arXiv:1803.06969*, 2018.

[13] M. Mézard, G. Parisi, and M. Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

[14] M. Mézard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[15] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[16] S. Rangan. Generalized approximate message passing for estimation with random linear mixing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2168–2172. IEEE, 2011.

[17] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[18] A. Javanmard and A. Montanari. State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Information and Inference: A Journal of the IMA*, 2(2):115–144, 2013.

[19] H. Schwarze. Learning a rule in a multilayer neural network. *Journal of Physics A: Mathematical and General*, 26(21):5781, 1993.

[20] H. Schwarze and J. Hertz. Generalization in a large committee machine. *EPL (Europhysics Letters)*, 20(4):375, 1992.

[21] H. Schwarze and J. Hertz. Generalization in fully connected committee machines. *EPL (Europhysics Letters)*, 21(7):785, 1993.

[22] G. Mato and N. Parga. Generalization properties of multilayered neural networks. *Journal of Physics A: Mathematical and General*, 25(19):5047, 1992.

[23] D. Saad and S. A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225, 1995.

[24] J. Barbier and N. Macris. The adaptive interpolation method: a simple scheme to prove replica formulas in bayesian inference. *Probability Theory and Related Fields*, pages 1–53, 2018.

[25] D. L. Donoho, I. Johnstone, and A. Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE transactions on information theory*, 59(6):3396–3433, 2013.

[26] L. Zdeborová and F. Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[27] Y. Deshpande and A. Montanari. Finding hidden cliques of size \sqrt {N/e} n/e in nearly linear time. *Foundations of Computational Mathematics*, 15(4):1069–1128, 2015.

[28] A. S. Bandeira, A. Perry, and A. S. Wein. Notes on computational-to-statistical gaps: predictions using statistical physics. *arXiv preprint arXiv:1803.11132*, 2018.

[29] I. Safran and O. Shamir. Spurious local minima are common in two-layer relu neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

[30] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. I. Jordan. Decoding from pooled data: Sharp information-theoretic bounds. *arXiv preprint arXiv:1611.09981*, 2016.

[31] A. El Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. I. Jordan. Decoding from pooled data: Phase transitions of message passing. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 2780–2784. IEEE, 2017.

[32] J. Zhu, D. Baron, and F. Krzakala. Performance limits for noisy multimeasurement vector problems. *IEEE Transactions on Signal Processing*, 65(9):2444–2454, 2017.

[33] F. Guerra. Broken replica symmetry bounds in the mean field spin glass model. *Communications in mathematical physics*, 233(1):1–12, 2003.

[34] M. Talagrand. *Spin glasses: a challenge for mathematicians: cavity and mean field models*, volume 46. Springer Science & Business Media, 2003.

[35] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of'solvable model of a spin glass'. *Philosophical Magazine*, 35(3):593–601, 1977.

[36] M. Mézard. The space of interactions in neural networks: Gardner's computation with the cavity method. *Journal of Physics A: Mathematical and General*, 22(12):2181–2190, 1989.

[37] M. Opper and O. Winther. Mean field approach to bayes learning in feed-forward neural networks. *Physical review letters*, 76(11):1964, 1996.

[38] Y. Kabashima. Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels. *Journal of Physics: Conference Series*, 95(1):012001, 2008.

[39] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina. Efficient supervised learning in networks with binary synapses. *Proceedings of the National Academy of Sciences*, 104(26):11079–11084, 2007.

[40] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová. AMP implementation of the committee machine. https://github.com/benjaminaubin/TheCommitteeMachine, 2018.

[41] P. Schniter, S. Rangan, and A. K. Fletcher. Vector approximate message passing for the generalized linear model. In *Signals, Systems and Computers, 2016 50th Asilomar Conference on*, pages 1525–1529. IEEE, 2016.

[42] G. Reeves, H. D. Pfister, and A. Dytso. Mutual information as a function of matrix snr for linear gaussian channels. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1754–1758. IEEE, 2018.

[43] M. Payaró, M. Gregori, and D. Palomar. Yet another entropy power inequality with an application. In *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, pages 1–5. IEEE, 2011.

[44] M. Lamarca. Linear precoding for mutual information maximization in mimo systems. In *Wireless Communication Systems, 2009. ISWCS 2009. 6th International Symposium on*, pages 26–30. IEEE, 2009.

[45] J. Barbier. Overlap matrix concentration in optimal bayesian inference. *arXiv preprint arXiv:1904.02808*, 2019.

[46] J. Barbier and N. Macris. The adaptive interpolation method for proving replica formulas. applications to the curie-weiss and wigner spike models. *Journal of Physics A: Mathematical and Theoretical*, 2019.

[47] J. Barbier, C. Luneau, and N. Macris. Mutual information for low-rank even-order symmetric tensor factorization. *arXiv preprint arXiv:1904.04565*, 2019.

[48] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08009, 2012.

[49] U. Kamilov, S. Rangan, M. Unser, and A. K. Fletcher. Approximate message passing with consistent parameter estimation and applications to sparse learning. In *Advances in Neural Information Processing Systems*, pages 2438–2446, 2012.

[50] P. Hartman. *Ordinary Differential Equations: Second Edition.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1982.

[51] E. Gardner and B. Derrida. Optimal storage properties of neural network models. *Journal of Physics A: Mathematical and general*, 21(1):271, 1988.

[52] J. Barbier, N. Macris, M. Dia, and F. Krzakala. Mutual information and optimality of approximate message-passing in random linear estimation. *arXiv preprint arXiv:1701.05823*, 2017.

[53] M. Opper and W. Kinzel. Statistical mechanics of generalization. In *Models of neural networks III*, pages 151–209. Springer, 1996.

[54] J. Barbier and F. Krzakala. Approximate message-passing decoder and capacity achieving sparse superposition codes. *IEEE Transactions on Information Theory*, 63:4894–4927, 2017.

[55] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[56] M. Bayati, M. Lelarge, A. Montanari, et al. Universality in polytope phase transitions and message passing algorithms. *The Annals of Applied Probability*, 25(2):753–822, 2015.

Supplementary material

A Proof details for Theorem 3.1

A.1 The Nishimori property in Bayes-optimal learning

We first state an important property of the Bayesian optimal setting (that is when all hyper-parameters of the problem are assumed to be known), that is used several times, and is often refered to as the Nishimori identity.

Proposition A.1 (Nishimori identity). *Let $(X, Y) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ be a couple of random variables. Let $k \geq 1$ and let $X^{(1)}, \dots, X^{(k)}$ be k i.i.d. samples (given Y) from the conditional distribution $P(X = \cdot\,|Y)$, independently of every other random variables. Let us denote $\langle - \rangle$ the expectation operator w.r.t. $P(X = \cdot\,|Y)$ and $\mathbb{E}$ the expectation w.r.t. (X, Y). Then, for all continuous bounded function g we have*

$$\mathbb{E}\langle g(Y, X^{(1)}, \dots, X^{(k)})\rangle = \mathbb{E}\langle g(Y, X^{(1)}, \dots, X^{(k-1)}, X)\rangle\,. \tag{31}$$

Proof. This is a simple consequence of Bayes formula. It is equivalent to sample the couple (X, Y) according to its joint distribution or to sample first Y according to its marginal distribution and then to sample X conditionally to Y from its conditional distribution $P(X = \cdot\,|Y)$. Thus the $(k+1)$-tuple $(Y, X^{(1)}, \dots, X^{(k)})$ is equal in law to $(Y, X^{(1)}, \dots, X^{(k-1)}, X)$. This proves the proposition. □

As a first application of Proposition A.1 we prove the following Lemma which is used in the proof of the upper bound Proposition 5.7.

Lemma A.2 (Positivity of some matrices). *The matrices ρ, $\mathbb{E}\langle Q\rangle$ and $\rho - \mathbb{E}\langle Q\rangle$ are positive definite, i.e. in $\mathcal{S}_K^+$. In the application the Gibbs bracket is $\langle - \rangle_{n,t,\epsilon}$.*

Proof. The statement for ρ follows from its definition (in Theorem 3.1). Note for further use that we also have $\rho = \frac{1}{n}\mathbb{E}[\underline{W}_i^*(\underline{W}_i^*)^\intercal]$. Since by definition $Q_{ll'} \equiv \frac{1}{n}\sum_{i=1}^n W_{il}^* w_{il'}$ in matrix notation we have $Q = \frac{1}{n}\sum_{i=1}^n \underline{W}_i^* \underline{w}_i^\intercal$. An application of the Nishimori identity shows that

$$\mathbb{E}\langle Q\rangle = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \underline{W}_i^* \underline{w}_i^\intercal\rangle = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\langle \underline{w}_i\rangle\langle \underline{w}_i^\intercal\rangle] \tag{32}$$

which is obviously in $\mathcal{S}_K^+$. Finally we note that

$$\mathbb{E}[\rho - \langle Q\rangle] = \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}[\underline{W}_i^*(\underline{W}_i^*)^\intercal] - \mathbb{E}[\langle \underline{w}_i\rangle\langle \underline{w}_i^\intercal\rangle]\Big) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\underline{W}_i^* - \langle \underline{w}_i\rangle)((\underline{W}_i^*)^\intercal - \langle \underline{w}_i^\intercal\rangle)]$$

where the last equality is proved by an application of the Nishimori identity again. This last expression is obviously in $\mathcal{S}_K^+$, i.e. $\mathbb{E}\langle Q\rangle \in \mathcal{S}_K^+(\rho)$. □

A.2 Setting in the Hamiltonian language

We set up some notations which will shortly be useful. Let $u_y(\underline{x}) \equiv \ln P_{\text{out}}(y|\underline{x})$. Here $\underline{x} \in \mathbb{R}^K$ and $y \in \mathbb{R}$. We will denote by $\nabla u_y(\underline{x})$ the K-dimensional gradient w.r.t. $\underline{x}$, and $\nabla\nabla^\intercal u_y(\underline{x})$ the $K \times K$ matrix of second derivatives (the Hessian) w.r.t. $\underline{x}$. Moreover $\nabla P_{\text{out}}(y|\underline{x})$ and $\nabla\nabla^\intercal P_{\text{out}}(y|\underline{x})$ also denote the K-dimensional gradient and Hessian w.r.t. $\underline{x}$. We will also use the matrix identity

$$\nabla\nabla^\intercal u_{Y_\mu}(\underline{x}) + \nabla u_{Y_\mu}(\underline{x})\nabla^\intercal u_{Y_\mu}(\underline{x}) = \frac{\nabla\nabla^\intercal P_{\text{out}}(Y_\mu|\underline{x})}{P_{\text{out}}(Y_\mu|\underline{x})}\,. \tag{33}$$

Finally we will use the matrices $w \in \mathbb{R}^{n\times K}$, $u \in \mathbb{R}^{m\times K}$, $Y_t \in \mathbb{R}^m$, $Y'_t \in \mathbb{R}^{n\times K}$, $X \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{m\times K}$, $W^* \in \mathbb{R}^{n\times K}$ and $U^* \in \mathbb{R}^{m\times K}$. Like in sec. 5 we adopt the convention that all underlined vectors are K-dimensional, like e.g. $\underline{u}_\mu$, $\underline{U}_\mu$, $\underline{V}_\mu$ and $\underline{Y}'_{t,i}$.

It is convenient to reformulate the expression of the interpolating free entropy $f_{n,\epsilon}(t)$ in the Hamiltonian language. We introduce an *interpolating Hamiltonian*:

$$\mathcal{H}_t(w,u;Y_t,Y'_t,X,V) \equiv -\sum_{\mu=1}^m u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) + \frac{1}{2}\sum_{i=1}^n \|\underline{Y}'_{t,i} - R_1(t)^{1/2}\,\underline{w}_i\|_2^2 \tag{34}$$

where recall that

$$\underline{s}_{t,\mu} \equiv \sqrt{\frac{1-t}{n}}\sum_{i=1}^n X_{\mu i}\underline{w}_i + \sqrt{R_2(t)}\,\underline{V}_\mu + \sqrt{t\rho - R_2(t) + 2s_n I_{K\times K}}\,\underline{u}_\mu\,. \tag{35}$$

The expression of $\mathcal{H}_t(W^*,U^*;Y_t,Y'_t,X,V)$ is similar to (34), but with w replaced by W^* and $\underline{s}_{t,\mu}$ given by (35) replaced by $\underline{S}_{t,\mu}$ given by (15). The average free entropy (18) at time t then reads

$$f_{n,\epsilon}(t) \equiv \frac{1}{n}\mathbb{E}\ln\int_{\mathbb{R}^{n\times K}} dP_0(w)\int_{\mathbb{R}^{m\times K}} \mathcal{D}u\, e^{-\mathcal{H}_t(w,u;Y_t,Y'_t,X,V)} \tag{36}$$

where $\mathcal{D}u = \prod_{\mu=1}^m\prod_{l=1}^K (2\pi)^{-1/2}e^{-u_{\mu l}^2/2}$ and $dP_0(w) = \prod_{i=1}^n P_0(\underline{w}_i)\prod_{l=1}^K dw_{il}$. To develop the calculations in the simplest manner it is fruitful to represent the expectations over W^*, U, Y, Y' explicitly as integrals:

$$f_{n,\epsilon}(t) = \frac{1}{n}\mathbb{E}_{X,V}\int dY_t dY'_t dP_0(W^*)\mathcal{D}U^* e^{-\mathcal{H}_t(W^*,U;Y_t,Y'_t,X,V)}\ln\int dP_0(w)\mathcal{D}u\, e^{-\mathcal{H}_t(w,u;Y_t,Y'_t,X,V)}. \tag{37}$$

A.3 Free entropy variation: Proof of Proposition 5.2

The proof provided here follows very closely the one in [11] for the case $K = 1$, so we are more brief and refer to this paper for more details. We first prove that for all $t \in (0,1)$

$$\begin{aligned}\frac{df_{n,\epsilon}(t)}{dt} =& -\frac{1}{2}\mathbb{E}\Big\langle \mathrm{Tr}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^m \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - r(t)\Big)\Big(\frac{1}{n}\sum_{i=1}^n \underline{W}_i^*\underline{w}_i^\intercal - q(t)\Big)\Big\rangle_{n,t,\epsilon}\\ &+\frac{1}{2}\mathrm{Tr}[r(t)(q(t)-\rho)] - \frac{A_n}{2}\,,\end{aligned} \tag{38}$$

where

$$A_n = \mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{\nabla\nabla^\intercal P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{t,\mu})}{P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{t,\mu})}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\underline{W}_i^*(\underline{W}_i^*)^\intercal - \rho)\Big)\Big]\frac{1}{n}\ln\mathcal{Z}_{n,\epsilon}(t)\Big]\,. \tag{39}$$

Once this is done, we show that A_n goes to 0 as $n\to\infty$ uniformly in $t\in[0,1]$ in order to conclude the proof.

The Hamiltonian (34) t-derivative evaluated at the ground-truth matrices is given by

$$\begin{aligned}\frac{d\mathcal{H}_t}{dt}(W^*,U^*;Y_t,Y'_t,X,V) &= -\sum_{\mu=1}^m \nabla^\intercal u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\frac{d\underline{S}_{t,\mu}}{dt} - \sum_{i=1}^n\Big(\frac{dR_1(t)^{1/2}}{dt}\underline{W}_i^*\Big)^\intercal(\underline{Y}'_{t,i} - R_1(t)^{1/2}\underline{W}_i^*)\\ &= -\sum_{\mu=1}^m \mathrm{Tr}\Big[\frac{d\underline{S}_{t,\mu}}{dt}\nabla^\intercal u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\Big] - \sum_{i=1}^n \mathrm{Tr}\Big[\Big(\frac{dR_1(t)^{1/2}}{dt}\Big)^\intercal(\underline{Y}'_{t,i} - R_1(t)^{1/2}\underline{W}_i^*)\underline{W}_i^{*\intercal}\Big]\end{aligned} \tag{40}$$

(where we used that $R_1(t)$ is symmetric). The t-derivative of $f_{n,\epsilon}(t)$ thus reads, for $0 < t < 1$,

$$\frac{df_{n,\epsilon}(t)}{dt} = -\underbrace{\frac{1}{n}\mathbb{E}\Big[\frac{d\mathcal{H}_t}{dt}(W^*, U^*; Y_t, Y_t', X, V)\ln \mathcal{Z}_{n,\epsilon}(t)\Big]}_{T_1} - \underbrace{\frac{1}{n}\mathbb{E}\Big\langle \frac{d\mathcal{H}_t}{dt}(w, u; Y_t, Y_t', X, V)\Big\rangle_{n,t,\epsilon}}_{T_2}. \tag{41}$$

First, we note that $T_2 = 0$. This is a direct consequence of the Nishimori identity Proposition A.1:

$$T_2 = \frac{1}{n}\mathbb{E}\Big\langle \frac{d\mathcal{H}_t}{dt}(w, u; Y_t, Y_t', X, V)\Big\rangle_{n,t,\epsilon} = \frac{1}{n}\mathbb{E}\,\frac{d\mathcal{H}_t}{dt}(W^*, U^*; Y_t, Y_t', X, V) = 0\,. \tag{42}$$

We now compute T_1. Starting from (40) and considering the first term only (recall also the expression (15) for $\underline{S}_{t,\mu}$),

$$\begin{aligned}\mathbb{E}\Big[\mathrm{Tr}\Big[\frac{d\underline{S}_{t,\mu}}{dt}\nabla^\intercal u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big] &= \mathbb{E}\Big[\mathrm{Tr}\Big[\Big\{-\frac{\sum_{i=1}^n X_{\mu i}\underline{W}_i^*}{2\sqrt{n(1-t)}}\\ &+ \frac{d}{dt}\sqrt{R_2(t)}\underline{V}_\mu + \frac{d}{dt}\sqrt{t\rho - R_2(t) + 2s_n I_{K\times K}}\,\underline{U}_\mu^*\Big\}\nabla^\intercal u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big]. \end{aligned}\tag{43}$$

We then compute the first line of the right-hand side of (43). By Gaussian integration by parts w.r.t. $X_{\mu i}$ (recall hypothesis (H3)), and using the identity (33), we find after some algebra

$$\begin{aligned}&-\frac{1}{2\sqrt{n(1-t)}}\mathbb{E}\Big[\mathrm{Tr}\Big[\sum_{i=1}^n X_{\mu i}\underline{W}_i^*\nabla^\intercal u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big]\\ &\quad= -\frac{1}{2}\mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{n}\sum_{i=1}^n \underline{W}_i^*\underline{W}_i^\intercal\frac{\nabla\nabla^\intercal P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{t,\mu})}{P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{t,\mu})}\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big]\\ &\qquad- \frac{1}{2}\mathbb{E}\Big\langle \mathrm{Tr}\Big[\frac{1}{n}\sum_{i=1}^n \underline{W}_i^*\underline{w}_i^\intercal\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})\nabla^\intercal u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\Big]\Big\rangle_{n,t,\epsilon}. \end{aligned}\tag{44}$$

Similarly for the second line of the right hand side of (43), we use again Gaussian integrations by parts but this time w.r.t. $\underline{V}_\mu, \underline{U}_\mu^*$ which have i.i.d. $\mathcal{N}(0,1)$ entries. This calculation has to be done carefully with the help of the matrix identity

$$\frac{d}{dt}M(t) = \sqrt{M(t)}\,\frac{d\sqrt{M(t)}}{dt} + \frac{d\sqrt{M(t)}}{dt}\sqrt{M(t)} \tag{45}$$

for any $M(t) \in \mathcal{S}_K^+$, and the cyclicity and linearity of the trace. Applying (45) to $M(t)$ equal to $\int_0^t q(s)ds$ and $\int_0^t(\rho - q(s))ds$, as well as the identity (33), we reach after some algebra

$$\begin{aligned}&\mathbb{E}\Big[\mathrm{Tr}\Big[\Big(\frac{d}{dt}\sqrt{R_2(t)}\underline{V}_\mu + \frac{d}{dt}\sqrt{t\rho - R_2(t) + 2s_n I_{K\times K}}\,\underline{U}_\mu^*\Big)\nabla^\intercal u_{Y_\mu}(\underline{S}_{\mu,t})\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big]\\ &= \mathbb{E}\Big[\mathrm{Tr}\Big[\rho\frac{\nabla\nabla^\intercal P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{\mu,t})}{P_{\mathrm{out}}(Y_{t,\mu}|\underline{S}_{\mu,t})}\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big] + \mathbb{E}\Big\langle \mathrm{Tr}\Big[q(t)\nabla u_{Y_{t,\mu}}(\underline{S}_{\mu,t})\nabla^\intercal u_{Y_{t,\mu}}(\underline{s}_{\mu,t})\Big]\Big\rangle_{n,t,\epsilon}. \end{aligned}\tag{46}$$

As seen from (40), (41) it remains to compute $\mathbb{E}[\mathrm{Tr}[(\frac{d}{dt}\sqrt{R_1(t)})^\intercal(\underline{Y}_{t,i}' - \sqrt{R_1(t)}\underline{W}_i^*)\underline{W}_i^{*\intercal}]\ln \mathcal{Z}_{n,\epsilon}(t)]$. Recall that $\underline{Y}_{t,i}' - \sqrt{R_1(t)}\underline{W}_i^* = \underline{Z}_i' \sim \mathcal{N}(0, I_{K\times K})$. Using Gaussian integration by parts as well as the identity (45) one obtains

$$\mathbb{E}\Big[\mathrm{Tr}\Big[\Big(\frac{d}{dt}\sqrt{R_1(t)}\Big)^\intercal(\underline{Y}_{t,i}' - \sqrt{R_1(t)}\underline{W}_i^*)\underline{W}_i^{*\intercal}\Big]\ln \mathcal{Z}_{n,\epsilon}(t)\Big] = -\mathrm{Tr}\Big[\sqrt{R_1(t)}\big(\rho - \mathbb{E}\langle W_j^* w_j\rangle_{n,t,\epsilon}\big)\Big]. \tag{47}$$

Finaly the term T_1 is obtained by putting together (43), (44), (46) and (47).

It now remains to check that $A_n \to 0$ as $n \to +\infty$ uniformly in $t \in [0,1]$. The proof from [11] (Appendix C.2) can easily be adapted so we give here just a few indications for the ease of the reader. First one notices that

$$\mathbb{E}\Big[\frac{\nabla\nabla^\intercal P_{\rm out}(Y_{t,\mu}|\underline{S}_{t,\mu})}{P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}\,\Big|\,W^*,\{\underline{S}_{t,\mu}\}_{\mu=1}^m\Big] = \int dY_\mu \nabla\nabla^\intercal P_{\rm out}(Y_{t,\mu}|\underline{S}_{t,\mu}) = 0\,, \tag{48}$$

so that by the tower property of the conditional expectation one gets

$$\mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{\nabla\nabla^\intercal P_{\rm out}(Y_{t,\mu}|\underline{S}_{t,\mu})}{P_{\rm out}(Y_{t,\mu}|\underline{S}_{t,\mu})}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\underline{W}_i^*(\underline{W}_i^*)^\intercal-\rho)\Big)\Big]\Big] = 0\,. \tag{49}$$

Next, one shows by standard second moment methods that $\mathbb{E}[(\ln \mathcal{Z}_{n,\epsilon}(t)/n - f_{n,\epsilon}(t))^2] \to 0$ as $n \to +\infty$ uniformly in $t \in [0,1]$ (see [11] for the proof at $K=1$, that generalizes straightforwardly for any finite K). Then, using this last fact together with (49), and under hypotheses (H1), (H2), (H3), an easy application of the Cauchy-Schwarz inequality implies $A_n \to 0$ as $n \to +\infty$ uniformly in $t \in [0,1]$. This ends the proof. □

A.4 Technical lemmas

Lemma A.3 (Cauchy-Lipschitz Theorem and Liouville Formula). *Let*

$$F : \left|\begin{array}{ccc} [0,1]\times(0,+\infty)^d & \to & [0,+\infty)^d \\ (t,z) & \mapsto & F(t,z)\end{array}\right.$$

be a continuous, bounded function. Assume that F admits continuous partial derivatives $\frac{\partial F}{\partial z_i}$ $(i=1,\dots,d)$ on its domain of definition. Then, for all $\epsilon \in (0,+\infty)^d$, the Cauchy problem

$$y(0) = \epsilon \qquad \text{and} \qquad y'(t) = F\big(t, y(t)\big) \tag{50}$$

admits a unique solution $t \mapsto y(t,\epsilon)$. For all $t \in [0,1]$, the mapping $z_t : \epsilon \mapsto y(t,\epsilon)$ is a diffeomorphism of class $\mathcal{C}^1$, from $(0,+\infty)^d$ to $z_t\big((0,+\infty)^d\big)$. Moreover the determinant $J(z_t)(\epsilon)$ of the Jacobian of z_t at ϵ verifies

$$J(z_t)(\epsilon) = \det\Big(\Big(\frac{\partial y_i}{\partial \epsilon_j}\Big)_{i,j}\Big) = \exp\Big(\int_0^t \sum_{i=1}^d \frac{\partial F_i}{\partial z_i}\big(s, y(s,\epsilon)\big)ds\Big)\,. \tag{51}$$

Thus, in particular, if in addition $\sum_{i=1}^d \frac{\partial F_i}{\partial z_i} \geq 0$ then $J(z_t)(\epsilon) \geq 1$ for all ϵ.

Proof. The existence and uniqueness of the solution of (50) follows from the classical Cauchy-Lipschitz Theorem. The solution is indeed defined on all the segment $[0,1]$ because F is bounded.

Theorem 3.1 from Chapter 5 in [50] gives that y admits continuous partial derivatives $\frac{\partial y}{\partial \epsilon_i}$ for $i=1,\dots,d$, and Corollary 3.1 from Chapter 5 in the same reference states the Liouville formula (51).

By the Cauchy-Lipschitz Theorem, two solutions of $y'(t) = F\big(t,y(t)\big)$ that are equal at some $t \in [0,1]$ are equal everywhere. This implies that the mapping $z_t : \epsilon \mapsto y(t,\epsilon)$ is injective, for all $t \in [0,1]$. Since y admits continuous partial derivatives in ϵ_i, $i=1,\dots,d$, we obtain that z_t is of class $\mathcal{C}^1$ on $(0,+\infty)^d$. Now, the equation (51) gives that $J(z_t)(\epsilon) > 0$ for all $\epsilon \in (0,+\infty)^d$. The local inversion Theorem gives then that z_t is a $\mathcal{C}^1$ diffeomorphism. □

Lemma A.4 (Boundedness of an overlap fluctuation). *Under hypothesis (H2) one can find a constant* $C(\varphi, K, \Delta) < +\infty$ *(independent of* n, t, ϵ*) such that for any* $R_n \in \mathcal{S}_K^+$ *we have*

$$\mathbb{E}\Big\langle \Big\| \frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - R_n \Big\|_{\mathrm{F}}^2 \Big\rangle_{n,t,\epsilon} \leq 2\mathrm{Tr}(R_n^2) + \alpha^2 C(\varphi, K, \Delta). \tag{52}$$

We note that the constant remains bounded as $\Delta \to 0$ *and diverges as* $K \to +\infty$.

Proof. It is easy to see that for symmetric matrices A, B we have $\mathrm{Tr}(A-B)^2 \leq 2(\mathrm{Tr}A^2 + \mathrm{Tr}B^2)$. Therefore

$$\begin{aligned}\mathbb{E}\Big\langle \Big\| \frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - R_n \Big\|_{\mathrm{F}}^2 \Big\rangle_{n,t,\epsilon} \\ \leq 2\mathrm{Tr}(R_n^2) + 2\mathbb{E}\Big\langle \mathrm{Tr}\Big(\frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal \Big)^2 \Big\rangle_{n,t,\epsilon}.\end{aligned} \tag{53}$$

In the rest of the argument we bound the second term of the r.h.s. Using the triangle inequality and then Cauchy-Schwarz we obtain

$$\begin{aligned}\mathbb{E}\Big\langle \Big\| \frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal \Big\|_{\mathrm{F}}^2 \Big\rangle_{n,t,\epsilon} &\leq \mathbb{E}\Big\langle \frac{1}{n^2}\Big(\sum_{\mu=1}^{m} \|\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal\|_{\mathrm{F}} \Big)^2 \Big\rangle_{n,t,\epsilon} \\ &\leq \mathbb{E}\Big\langle \frac{1}{n^2}\Big(\sum_{\mu=1}^{m} \|\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\|_2 \|\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal\|_2 \Big)^2 \Big\rangle_{n,t,\epsilon}.\end{aligned} \tag{54}$$

From the random representation of the transition kernel,

$$u_{Y_{t,\mu}}(\underline{s}) = \ln P_{\mathrm{out}}(Y_{t,\mu}|\underline{x}) = \ln \int dP_A(a_\mu) \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}(Y_{t,\mu} - \varphi(\underline{x}, a_\mu))^2} \tag{55}$$

and thus

$$\nabla u_{Y_{t,\mu}}(\underline{x}) = \frac{\int dP_A(a_\mu)(Y_{t,\mu} - \varphi(\underline{x}, a_\mu)) \nabla\varphi(\underline{x}, a_\mu) e^{-\frac{1}{2\Delta}(Y_{t,\mu} - \varphi(\underline{x}, a_\mu))^2}}{\int dP_A(a_\mu) e^{-\frac{1}{2\Delta}(Y_{t,\mu} - \varphi(\underline{x}, a_\mu))^2}} \tag{56}$$

where $\nabla\varphi$ is the K-dimensional gradient w.r.t. the first argument $\underline{x} \in \mathbb{R}^K$. From the observation model we get $|Y_{t,\mu}| \leq \sup|\varphi| + \sqrt{\Delta}|Z_\mu|$, where the supremum is taken over both arguments of φ, and thus we immediately obtain for all $\underline{s} \in \mathbb{R}^K$

$$\|\nabla u_{Y_{t,\mu}}(\underline{x})\| \leq (2\sup|\varphi| + \sqrt{\Delta}|Z_\mu|) \sup\|\nabla\varphi\|. \tag{57}$$

From (57) and (54) we see that it suffices to check that

$$\frac{m^2}{n^2}\mathbb{E}\big[\big((2\sup|\varphi| + |Z_\mu|)^2 (\sup\|\nabla\varphi\|)^2\big)^2\big] \leq C(\varphi, K, \Delta)$$

where $C(\varphi, K, \Delta) < +\infty$ is a finite constant depending only on φ, K, and Δ. This is easily seen by expanding all squares and using that $m/n \to \alpha$. This ends the proof of Lemma A.4. □

Lemma A.5 (Properties of ψ_{P_0}). ψ_{P_0} *is defined as the free entropy of the first auxiliary channel* (3). *We have,*

for any $r \in \mathcal{S}_K^+$:

$$\psi_{P_0}(r) \equiv \mathbb{E} \ln \int_{\mathbb{R}^K} \mathrm{d}w P_0(w) e^{Y_0^\intercal r^{1/2} w - \frac{1}{2} w^\intercal r w}.$$

Then ψ_{P_0} is convex and differentiable on $\mathcal{S}_K^+$, with $\nabla \psi_{P_0}(r) \in \mathcal{S}_K^+$ for any $r \in \mathcal{S}_K^+$.

Proof. Note that ψ_{P_0} is related to the mutual information $I(W_0; Y_0)$ via the relation $I(W_0; Y_0) = -\psi_{P_0}(r) + \frac{K}{2} + \frac{1}{2}\mathrm{Tr}[r\rho]$. It is then a known result (see [42, 43, 44]) that the derivative $\nabla_r I(W_0; Y_0)$ is given by the matrix-MMSE, i.e. $\nabla_r I(W_0; Y_0) = \frac{1}{2}\mathbb{E}\left[\langle w \rangle \langle w \rangle^\intercal\right]$. This implies that $\nabla_r \psi_{P_0}(r) = \frac{1}{2}(\rho - \mathbb{E}[\langle w \rangle \langle w \rangle^\intercal])$. Using the Nishimori identity Prop.A.1, we can write it as $\nabla_r \psi_{P_0}(r) = \frac{1}{2}\mathbb{E}\left[(w - \langle w \rangle)(w - \langle w \rangle)^\intercal\right]$, which is clearly a positive matrix. It is also known (see for instance Lemma 4 of [42]), that $I(W_0; Y_0)$ is a concave function of r, which implies that ψ_{P_0} is convex, which ends the proof. □

Lemma A.6 (Properties of $\Psi_{P_{\mathrm{out}}}$). *Recall that $\Psi_{P_{\mathrm{out}}}$ is defined as the free entropy of the second auxiliary channel (4). More precisely, for $q \in \mathcal{S}_K^+(\rho)$, we have:*

$$\Psi_{P_{\mathrm{out}}}(q) \equiv \mathbb{E} \ln \int_{\mathbb{R}^K} dw \frac{e^{-\frac{1}{2}\|w\|^2}}{(2\pi)^{K/2}} P_{\mathrm{out}}\big(\widetilde{Y}_0 | q^{1/2} V + (\rho - q)^{1/2} w\big).$$

Then $\Psi_{P_{\mathrm{out}}}$ is continuous and convex on $\mathcal{S}_K^+(\rho)$, and twice differentiable inside $\mathcal{S}_K^+(\rho)$. Also, $\nabla \Psi_{P_{\mathrm{out}}}(q) \in \mathcal{S}_K^+$.

Proof. The continuity and differentiability of $\Psi_{P_{\mathrm{out}}}$ is easy, and exactly similar to the first part of the proof of Proposition 18 of [11]; it just follows from the hypothesis (H2) which allows to use continuity and differentiation under the expectation, because all the domination hypotheses are easily verified.

One can compute the gradient and Hessian matrix of $\Psi_{P_{\mathrm{out}}}(q)$, for q inside $\mathcal{S}_K^+(\rho)$, using Gaussian integration by parts and the Nishimori identity. The calculation is tedious and essentially follows the steps of Proposition 11 of [11]. Recall that $u_{\widetilde{Y}_0}(x) \equiv \ln P_{\mathrm{out}}(\widetilde{Y}_0 | x)$. We define the average $\langle - \rangle_{\mathrm{sc}}$ (where sc stands for "scalar channel") as

$$\langle g(w) \rangle_{\mathrm{sc}} \equiv \frac{\int_{\mathbb{R}^K} \mathcal{D}w P_{\mathrm{out}}(\widetilde{Y}_0 | (\rho - q)^{1/2} w + q^{1/2} V) g(w)}{\int_{\mathbb{R}^K} \mathcal{D}w P_{\mathrm{out}}(\widetilde{Y}_0 | (\rho - q)^{1/2} w + q^{1/2} V)}, \tag{58}$$

for any continuous bounded function g. One arrives at:

$$\nabla \Psi_{P_{\mathrm{out}}}(q) = \frac{1}{2}\mathbb{E}\Big\langle \nabla u_{\widetilde{Y}_0}\Big((\rho - q)^{1/2} W^* + q^{1/2} V\Big) \nabla u_{\widetilde{Y}_0}\Big((\rho - q)^{1/2} w + q^{1/2} V\Big)^\intercal \Big\rangle_{\mathrm{sc}}. \tag{59}$$

Note that this gradient is actually a symmetric matrix of size $K \times K$, as it is a gradient w.r.t. q, which is itself a matrix of size K. The Hessian $\nabla \nabla^\intercal \Psi_{P_{\mathrm{out}}}$ with respect to q is thus a 4-tensor. One can compute in the same way:

$$\begin{aligned} \nabla \nabla^\intercal \Psi_{P_{\mathrm{out}}}(q) = \frac{1}{2}\mathbb{E}\Big[\Big(\Big\langle \frac{\nabla \nabla^\intercal P_{\mathrm{out}}(\widetilde{Y}_0 | (\rho - q)^{1/2} w + q^{1/2} V)}{P_{\mathrm{out}}(\widetilde{Y}_0 | (\rho - q)^{1/2} w + q^{1/2} V)} \Big\rangle_{\mathrm{sc}} \\ - \Big\langle \nabla u_{\widetilde{Y}_0}\Big((\rho - q)^{1/2} W^* + q^{1/2} V\Big) \nabla u_{\widetilde{Y}_0}\Big((\rho - q)^{1/2} w + q^{1/2} V\Big)^\intercal \Big\rangle_{\mathrm{sc}}\Big)^{\otimes 2}\Big]. \end{aligned} \tag{60}$$

In this expression, $\otimes 2$ means the "tensorized square" of a matrix, i.e. for any matrix M of size $K \times K$, $M^{\otimes 2}$ is a 4-tensor with indices $M^{\otimes 2}_{l_0 l_1 l_2 l_3} = M_{l_0 l_1} M_{l_2 l_3}$. From this expression, it is clear that the Hessian of $\Psi_{P_{\mathrm{out}}}$ is always positive, when seen as a matrix with rows and columns in $\mathcal{S}_K$, and thus $\Psi_{P_{\mathrm{out}}}$ is convex, which ends the proof of Lemma A.6. □

B Replica calculation

Our goal here is to provide an heuristic derivation of the replica formula of Theorem 3.1 using the replica method, a powerful non-rigorous tool from statistical physics of disordered systems [13, 14]. This computation is necessary to properly "guess" the formula that we then prove using the adaptive interpolation method. The reader interested in the replica approach to neural networks and the commitee machine is invited to look as well to some of the classical papers [51, 36, 20, 21, 19, 5].

The replica trick makes use of the formula, for a random variable $x \in \mathbb{R}^n$ and a strictly positive function $f_n : \mathbb{R}^n \to \mathbb{R}$ that depends on n:

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E} \ln f_n = \lim_{p\to 0^+} \lim_{n\to\infty} \frac{1}{np} \ln \mathbb{E} f_n^p. \tag{61}$$

Note that the inversion of the two limits here is non-rigorous. Computing the moments $\mathbb{E} f^p$ can often be done for integers $p \in \mathbb{N}$, and one can conjecture from it its value for every $p > 0$, before taking the limit $p \to 0^+$ in (61) by analytical continuation of the value for integer p.

In our calculation, we will use this formula to compute the *free entropy* of our system, $f \equiv \lim_{n\to\infty} f_n$. We will thus need the moments of the partition function, for integer p:

$$\begin{aligned}
\mathbb{E}\mathcal{Z}_n^p &= \mathbb{E}\left[\int_{\mathbb{R}^n\times\mathbb{R}^K} dw \prod_{i=1}^n P_0\left(\{w_{il}\}_{l=1}^K\right) \prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu \middle| \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i} w_{il}\right\}_{l=1}^K\right)\right]^p, \\
&= \mathbb{E}\left[\prod_{a=1}^p \int_{\mathbb{R}^n\times\mathbb{R}^K} dw^a \prod_{i=1}^n P_0\left(\{w_{il}^a\}_{l=1}^K\right) \prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu \middle| \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i} w_{il}^a\right\}_{l=1}^K\right)\right].
\end{aligned}$$

The outer expectation is done over $X_{\mu i} \sim \mathcal{N}(0,1)$, $w^\star$ and Y. Writing $w^\star$ as w^0 we have:

$$\begin{aligned}
\mathbb{E}\mathcal{Z}_n^p = \mathbb{E}_X \int_{\mathbb{R}^m} dY \prod_{a=0}^p \Bigg[&\int_{\mathbb{R}^n\times\mathbb{R}^K} dw^a \prod_{i=1}^n P_0\left(\{w_{il}^a\}_{l=1}^K\right) \\
&\times \prod_{\mu=1}^m P_{\text{out}}\left(Y_\mu \middle| \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i} w_{il}^a\right\}_{l=1}^K\right)\Bigg].
\end{aligned}$$

To perform the average over X, we notice that, since it is an i.i.d. standard Gaussian matrix, then for every a, μ, l, $Z_{\mu l}^a \equiv n^{-1/2}\sum_{i=1}^n X_{\mu i} w_{il}^a$ follows a Gaussian multivariate distribution, with zero mean. This naturally leads to introduce its covariance tensor, which is equal to:

$$\mathbb{E} Z_{\mu l}^a Z_{\nu l'}^b = \delta_{\mu\nu} \Sigma_{\substack{al \\ bl'}} = \delta_{\mu\nu} Q_{bl'}^{al}, \tag{62}$$

$$Q_{bl'}^{al} \equiv \frac{1}{n}\sum_{i=1}^n w_{il}^a w_{il'}^b. \tag{63}$$

For every a, b, $Q_b^a \in \mathbb{R}^{K\times K}$ is the *overlap* matrix, and Σ is of size size $(p+1)K \times (p+1)K$. Introducing δ functions for fixing Q, we arrive at :

$$\mathbb{E}\left[\mathcal{Z}_n^p\right] = \prod_{(a,r)} \int_{\mathbb{R}} dQ_{ar}^{ar} \prod_{\{(a,r);(b,r')\}} \int_{\mathbb{R}} dQ_{br'}^{ar} \left[I_{\text{prior}}(\{Q_{br'}^{ar}\}) \times I_{\text{channel}}(\{Q_{br'}^{ar}\})\right], \tag{64}$$

with:

$$I_{\text{prior}}(\{Q^{ar}_{br'}\}) = \prod_{a=0}^{p}\left[\int_{\mathbb{R}^{n\times K}} dw^a P_0(w^a)\right]\left[\prod_{\{(a,l);(b,l')\}}\delta\left(Q^{al}_{bl'} - \frac{1}{n}\sum_{i=1}^{n} w^a_{il}w^b_{il'}\right)\right], \tag{65}$$

$$I_{\text{channel}}(\{Q^{ar}_{br'}\}) = \int_{\mathbb{R}^m} dY \prod_{a=0}^{p}\int_{\mathbb{R}^{m\times K}} dZ^a \prod_{a=0}^{p} P_{\text{out}}(Y|Z^a)e^{-\frac{m}{2}\ln\det\Sigma - \frac{mK(p+1)}{2}\ln 2\pi}$$
$$\exp\left[-\frac{1}{2}\sum_{\mu=1}^{m}\sum_{a,b}\sum_{l,l'} Z^a_{\mu l}Z^b_{\mu l'}(\Sigma^{-1})^{al}_{bl'}\right]. \tag{66}$$

By Fourier expanding the delta functions in I_{prior}, and performing a saddle-point method, one obtains:

$$\lim_{n\to\infty}\frac{1}{n}\ln\mathbb{E}\left[\mathcal{Z}_n^p\right] = \text{extr}_{Q,\hat{Q}}\left[H(Q,\hat{Q})\right], \tag{67}$$

in which (recall $\alpha \equiv \lim_{n\to\infty} m/n$) :

$$H(Q,\hat{Q}) \equiv \frac{1}{2}\sum_{a=0}^{p}\sum_{l,l'} Q^{al}_{al}\hat{Q}^{al}_{al} - \frac{1}{2}\sum_{a\neq b}\sum_{l,l'} Q^{al}_{bl'}\hat{Q}^{al}_{bl'} + \ln I + \alpha\ln J, \tag{68}$$

in which we defined:

$$I \equiv \prod_{a=0}^{p}\int_{\mathbb{R}^K} dw^a P_0(w^a)\exp\left[-\frac{1}{2}\sum_{a=0}^{p}\sum_{l,l'}\hat{Q}^{al}_{al'}w^a_l w^a_{l'} + \frac{1}{2}\sum_{a\neq b}\sum_{l,l'}\hat{Q}^{al}_{bl'}w^a_l w^b_{l'}\right], \tag{69}$$

$$J \equiv \int_{\mathbb{R}} dy \prod_{a=0}^{p}\int_{\mathbb{R}^K}\frac{dZ^a}{(2\pi)^{K(p+1)/2}}\frac{P_{\text{out}}(y|Z^a)}{\sqrt{\det\Sigma}}\exp\left[-\frac{1}{2}\sum_{a,b=0}^{p}\sum_{l,l'=1}^{K} Z^a_l Z^b_{l'}(\Sigma^{-1})^{al}_{bl'}\right]. \tag{70}$$

Our goal is to express $H(Q,\hat{Q})$ as an analytical function of p, in order to perform the replica trick. To do so, we will assume that the extremum of H is attained at a point in $Q,\hat{Q}$ space such that a *replica symmetry* property is verified. More concretely, we assume:

$$\exists Q^0 \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall a \in [|0,p|] \quad \forall (l,l') \in [|1,K|]^2 \quad Q^{al}_{al'} = Q^0_{ll'}, \tag{71}$$

$$\exists q \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall (a<b) \in [|0,p|]^2 \quad \forall (l,l') \in [|1,K|]^2 \quad Q^{al}_{bl'} = q_{ll'}, \tag{72}$$

and samely for $\hat{Q}^0$ and $\hat{q}$. Note that Q^0 is by definition a symmetric matrix, while q is also symmetric by our assumption of replica symmetry. Under this ansatz, we obtain:

$$H(Q^0,\hat{Q}^0,q,\hat{q}) = \frac{p+1}{2}\text{Tr}[Q^0\hat{Q}^0] - \frac{p(p+1)}{2}\text{Tr}[q\hat{q}] + \ln I + \alpha\ln J. \tag{73}$$

Remains now to compute an expression for I and J that is analytical in p, in order to take the limit $p\to 0^+$. This can be done easily, using the identity, for any symmetric positive matrix $M\in\mathbb{R}^{K\times K}$ and any vector $x\in\mathbb{R}^K$: $\exp\left(x^\intercal(M/2)x\right) = \int_{\mathbb{R}^K}\mathcal{D}\xi\,\exp\left(\xi^\intercal M^{1/2}x\right)$, in which $\mathcal{D}\xi$ is the standard Gaussian measure on $\mathbb{R}^K$.

We obtain:

$$I = \int_{\mathbb{R}^K} \mathcal{D}\xi \left[\int_{\mathbb{R}^K} dw\, P_0(w)\, \exp\left[-\frac{1}{2} w^\intercal (\hat{Q}^0 + \hat{q}) w + \xi^\intercal \hat{q}^{1/2} w \right] \right]^{p+1}, \tag{74}$$

$$J = \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi \left[\int_{\mathbb{R}^K} dZ P_{\text{out}}\left\{ y | (Q^0 - q)^{1/2} Z + q^{1/2} \xi \right\} \right]^{p+1}. \tag{75}$$

Our assumptions must be consistent in the sense that $\text{extr}_{Q,\hat{Q}} \left[\lim_{p\to 0^+} H(Q, \hat{Q}) \right] = 0$ (because $\mathbb{E}\mathcal{Z}_n^0 = 1$). In the $p \to 0^+$ limit, one easily gets $J = 1$ and $I = \int_{\mathbb{R}^K} dw\, P_0(w) \exp\left[-\frac{1}{2} w^\intercal \hat{Q}^0 w^0 \right]$. This implies that the optimal overlap parameters satisfy $\hat{Q}^0 = 0$ and $Q^0_{ll'} = \mathbb{E}_{P_0}\left[w_l w_{l'} \right]$. In the end, we obtain the final formula for the free entropy:

$$\lim_{n\to\infty} f_n = \text{extr}_{q,\hat{q}} \left\{ -\frac{1}{2} \text{Tr}[q\hat{q}] + I_P + \alpha I_C \right\}, \tag{76}$$

$$I_P \equiv \int_{\mathbb{R}^K} \mathcal{D}\xi \int_{\mathbb{R}^K} dw^0 P_0(w^0) \exp\left[-\frac{1}{2} (w^0)^\intercal \hat{q} w^0 + \xi^\intercal \hat{q}^{1/2} w^0 \right] \times \ln \left[\int_{\mathbb{R}^K} dw P_0(w) \exp\left[-\frac{1}{2} w^\intercal \hat{q} w + \xi^\intercal \hat{q}^{1/2} w \right] \right],$$

$$I_C \equiv \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi \int_{\mathbb{R}^K} \mathcal{D}Z^0 P_{\text{out}}\left\{ y | (Q^0 - q)^{1/2} Z^0 + q^{1/2} \xi \right\} \times \ln \left[\int_{\mathbb{R}^K} \mathcal{D}Z P_{\text{out}}\left\{ y | (Q^0 - q)^{1/2} Z + q^{1/2} \xi \right\} \right].$$

A known ambiguity of the replica method is that its result is given as an extremum, here over the set $\mathcal{S}_K^+(Q_0)$ of positive symmetric matrices, such that $(Q^0 - q)$ is also a positive matrix. It is easy to show that this form gives back the form given in Theorem 3.1, by assuming that this extremum is realized as a $\sup_{\hat{q}} \inf_q$. Note that in the notations of Theorem 3.1, Q^0 is denoted ρ and $\hat{q}$ is denoted R.

C Generalization error

We detail here two different possible definitions of the generalization error, and how they are related in our system. Recall that we wish to estimate W^* from the observation of $\varphi_{\text{out}}(XW^*)$. In the following, we denote $\mathbb{E}$ for the average over the (quenched) W^* and the data X, and $\langle - \rangle$ for the Gibbs average over the posterior distribution of W. One can naturally define the *Gibbs generalization error* as:

$$\epsilon_g^{\text{Gibbs}} \equiv \frac{1}{2} \mathbb{E}_{W^*,X} \big\langle \left[\varphi_{\text{out}}(XW) - \varphi_{\text{out}}(XW^*) \right]^2 \big\rangle, \tag{77}$$

and define the *Bayes-optimal generalization error* as:

$$\epsilon_g^{\text{Bayes}} \equiv \frac{1}{2} \mathbb{E}_{W^*,X} \big[\big(\langle \varphi_{\text{out}}(XW) \rangle - \varphi_{\text{out}}(XW^*) \big)^2 \big]. \tag{78}$$

Using the Nishimori identity A.1, one can show that:

$$\begin{aligned}\epsilon_g^{\text{Bayes}} &= \frac{1}{2}\mathbb{E}_{X,W^*}\left[\varphi_{\text{out}}\left(XW^*\right)^2\right] + \frac{1}{2}\mathbb{E}_{X,W^*}\left[\left\langle\varphi_{\text{out}}\left(XW\right)\right\rangle^2\right] \\ &\qquad - \mathbb{E}_{X,W^*}\left\langle\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right\rangle, \\ &= \frac{1}{2}\mathbb{E}_{X,W^*}\left[\varphi_{\text{out}}\left(XW^*\right)^2\right] - \frac{1}{2}\mathbb{E}_{X,W^*}\left\langle\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right\rangle.\end{aligned}$$

Using again the Nishimori identity one can write:

$$\epsilon_g^{\text{Gibbs}} = \mathbb{E}_{X,W^*}\left[\varphi_{\text{out}}\left(XW^*\right)^2\right] - \mathbb{E}_{X,W^*}\left\langle\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right\rangle,$$

which shows that $\epsilon_g^{\text{Gibbs}} = 2\epsilon_g^{\text{Bayes}}$. Note finally that since the distribution of X is rotationally invariant, the quantity $\mathbb{E}_X\left[\varphi_{\text{out}}\left(XW^*\right)\varphi_{\text{out}}\left(XW\right)\right]$ only depends on the *overlap* $q \equiv W^\intercal W^*$. As the overlap is shown to concentrate under the Gibbs measure by Proposition 5.3, and as we expect that the value it concentrates on is the optimum q^* of the replica formula (such fact is proven, e.g., for random linear estimation problems in [52]), the generalization error can itself be evaluated as a function of q^*. Examples where it is done include [53, 3, 19, 11].

C.1 The generalization error at $K = 2$

In this subsection alone, we go back to the $K = 2$ case, instead of the $K \to \infty$ limit. From the definition of the generalization error (see sec. C), one can directly give an explicit expression of this error in the $K = 2$ case. Recall our committee-symmetric assumption on the overlap matrix, which here reads

$$q = \begin{pmatrix} q_d + \frac{q_a}{2} & \frac{q_a}{2} \\ \frac{q_a}{2} & q_d + \frac{q_a}{2} \end{pmatrix}.$$

For concision, we denote here $\text{sign}(x) = \sigma(x)$. One obtains from (78):

$$\begin{aligned}\frac{1}{2} - 2\epsilon_g^{\text{Bayes,K=2}} &= \int_{\mathbb{R}^4} \mathcal{D}x\, \sigma\left[\sigma(x_1) + \sigma(x_2)\right] \\ &\times \sigma\left\{\sigma\left[(\frac{q_a}{2}+q_d)x_1 + \frac{q_a}{2}x_2 + x_3\sqrt{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}\right]\right. \\ &\left.+\sigma\left[\frac{q_a}{2}x_1 + (\frac{q_a}{2}+q_d)x_2 - x_3\frac{q_a(q_d+\frac{q_a}{2})}{\sqrt{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}} + x_4\sqrt{\frac{(1-q_d^2)(1-(q_a+q_d)^2)}{1-\frac{q_a^2}{2}-q_aq_d-q_d^2}}\right]\right\}.\end{aligned} \tag{79}$$

Note that one could possibly simplify this expression by using an appropriate orthogonal transformation on x. These integrals were then computed using Monte-Carlo methods to obtain the generalization error in the left and middle plots of Fig. 2.

D The large K limit in the committee symmetric setting

We consider the large K limit[2] for a sign activation function, and for different priors on the weights. Since the output is a sign, the channel is simply a delta function. We assume a committee symmetric solution, i.e. the

[2] A similar limit has been derived in the context of coding with sparse superposition codes [54]. There the large input alphabet limit of the mutual information is considered *after* the thermodynamic limit $n \to \infty$ corresponding to the large codeword limit in this coding context.

matrices q and $\hat{q}$ (q and R in the notations of Theorem 3.1) are of the type $q = q_d \mathbb{1}_K + \frac{q_a}{K} \mathbf{1}_K \mathbf{1}_K^\intercal$, with the unit vector $\mathbf{1}_K = (1)_{l=1}^K$, and similarly for $\hat{q}$. In the large K limit, this scaling of the order parameters is natural. Indeed, assume that the covariance of the prior is $Q^0 = \mathbb{1}_K$ ($Q^0 = \rho$ in the notations of Theorem 3.1). Since both q and $(Q^0 - q)$ are assumed to be positive matrices, it is easily shown to imply that $0 \leq q_d \leq 1$ and $0 \leq q_a + q_d \leq 1$.

D.1 Large K limit for sign activation function

In the following, we consider $Q^0 = \sigma^2 \mathbb{1}_K$. We are interested here in computing the leading order term in I_C of (76). Note that replacing σ^2 by 1 in this equation only amounts to replacing q by q/σ^2, so we can assume $\sigma^2 = 1$ without loss of generality. We (abusively) write I_C in (76) as $I_C = \sum_{y=\pm 1} \int_{\mathbb{R}^K} \mathcal{D}\xi \, I_C(y, \xi) \log I_C(y, \xi)$, with the definition

$$I_C(y, \xi) \equiv \int_{\mathbb{R}^K} \mathcal{D}Z P_{\text{out}} \left\{ y | (Q^0 - q)^{1/2} Z + q^{1/2} \xi \right\}. \tag{80}$$

Here, we assumed a sign activation function and no noise, as well as a particular form for Q_0 and q (see the remarks above). Note that this implies that $q^{1/2} = \sqrt{q_d} \mathbb{1}_K + \frac{\sqrt{q_a + q_d} - \sqrt{q_d}}{K} \mathbf{1}_K \mathbf{1}_K^\intercal$ and that $(Q_0 - q)^{1/2} = \sqrt{1 - q_d} \mathbb{1}_K + \frac{\sqrt{1 - q_a - q_d} - \sqrt{1 - q_d}}{K} \mathbf{1}_K \mathbf{1}_K^\intercal$. All together, this gives the following explicit expression for $I_C(y, \xi)$:

$$I_C(y, \xi) \equiv \int_{\mathbb{R}^K} \mathcal{D}Z \\ \times \delta \left\{ y - \text{sign} \left[\frac{1}{\sqrt{K}} \sum_{l=1}^K \text{sign} \left[\sqrt{1 - q_d} Z_l + \left(\sqrt{1 - q_a - q_d} - \sqrt{1 - q_d} \right) \frac{\mathbf{1}_K^\intercal Z}{K} + (q^{1/2} \xi)_l \right] \right] \right\}.$$

Introducing a new variable $w \equiv \frac{\mathbf{1}_K^\intercal Z}{\sqrt{K}}$ and a Fourier-transform of the then-introduced delta function, as well as another variable u being the argument of the outer sign function in the previous equations, one obtains:

$$I_C(y, \xi) = \int_{\mathbb{R}} \frac{dw d\hat{w}}{2\pi} \frac{du d\hat{u}}{2\pi} e^{iw\hat{w} + iu\hat{u}} \delta_{y, \text{sign}(u)} \\ \times \prod_{l=1}^K \int_{\mathbb{R}} \mathcal{D}z e^{-i\hat{w} \frac{z}{\sqrt{K}}} e^{-\frac{i\hat{u}}{\sqrt{K}} \text{sign} \left[z + \left[\sqrt{\frac{1 - q_a - q_d}{1 - q_d}} - 1 \right] \frac{w}{\sqrt{K}} + \frac{1}{\sqrt{1 - q_d}} (q^{1/2} \xi)_l \right]}.$$

Denote

$$\lambda_l(w, \xi) \equiv \left[\sqrt{\frac{1 - q_a - q_d}{1 - q_d}} - 1 \right] \frac{w}{\sqrt{K}} + \frac{1}{\sqrt{1 - q_d}} (q^{1/2} \xi)_l,$$

such that

$$I_C(y, \xi) = \int_{\mathbb{R}} \frac{dw d\hat{w}}{2\pi} \frac{du d\hat{u}}{2\pi} e^{iw\hat{w} + iu\hat{u}} \delta_{y, \text{sign}(u)} \prod_{l=1}^K \int_{\mathbb{R}} \mathcal{D}z e^{-i\hat{w} \frac{z}{\sqrt{K}}} e^{-\frac{i\hat{u}}{\sqrt{K}} \text{sign}[z + \lambda_l(w, \xi)]}.$$

For $1 \leq l \leq K$, one can rewrite the factorized integral in the last expression of $I_C(y,\xi)$ as:

$$I_C(y,\xi) = \int_{\mathbb{R}} \frac{dwd\hat{w}}{2\pi} \frac{dud\hat{u}}{2\pi} e^{iw\hat{w}+iu\hat{u}} \delta_{y,\text{sign}(u)} \prod_{l=1}^{K} J\left(\lambda_l(w,\xi),\hat{w},\hat{u}\right), \tag{81}$$

$$J\left(\lambda_l(w,\xi),\hat{w},\hat{u}\right) \equiv e^{-\frac{\lambda_l^2}{2}+i\lambda_l\frac{\hat{w}}{\sqrt{K}}} \int_{\mathbb{R}} \mathcal{D}z e^{z(\lambda_l - i\frac{\hat{w}}{\sqrt{K}})} e^{-\frac{i\hat{u}}{\sqrt{K}}\text{sign}[z]}. \tag{82}$$

We abusively dropped the dependency of λ_l on (w,ξ). Note the following identity:

$$F(\alpha,i\beta) \equiv \int_{\mathbb{R}} \mathcal{D}z e^{\alpha z + i\beta\ \text{sign}(z)} = e^{\alpha^2/2}\left[\cos\beta + i\sin\beta \hat{H}(\alpha)\right], \tag{83}$$

with $\hat{H}(x) = \text{erf}(x/\sqrt{2})$. Using it in our previous expressions, we obtain:

$$J(\lambda_l,\hat{w},\hat{u}) = e^{-\frac{1}{2K}\hat{w}^2}\left[\cos\left(\frac{\hat{u}}{\sqrt{K}}\right) - i\sin\left(\frac{\hat{u}}{\sqrt{K}}\right)\hat{H}\left(\lambda_l - i\frac{\hat{w}}{\sqrt{K}}\right)\right].$$

Note that by our committee-symmetry assumption, we have $\lambda_l(w,\xi) = \lambda_{l,0}(\xi) + \frac{1}{\sqrt{K}}\lambda_1(w,\xi)$ with $\lambda_{l,0}$ and λ_1 typically of order 1 when $K \to \infty$:

$$\lambda_{l,0}(\xi) \equiv \sqrt{\frac{q_d}{1-q_d}}\xi_l, \tag{84}$$

$$\lambda_1(w,\xi) \equiv \left[\sqrt{\frac{1-q_a-q_d}{1-q_d}} - 1\right] w + \left[\sqrt{\frac{q_a+q_d}{1-q_d}} - \sqrt{\frac{q_d}{1-q_d}}\right] \frac{\mathbf{1}_K^\intercal \xi}{\sqrt{K}}. \tag{85}$$

Expanding $J(\lambda_l,\hat{w},\hat{u})$ as $K \to \infty$, we obtain using the known development of the error function:

$$J(\lambda_l,\hat{w},\hat{u}) = e^{-\frac{1}{2K}\hat{w}^2}\left[1 - \frac{\hat{u}^2}{2K} - i\hat{H}\left[\lambda_{l,0}(\xi)\right]\frac{\hat{u}}{\sqrt{K}} - i\frac{\hat{u}\left[\lambda_1(w,\xi) - i\hat{w}\right]}{K}\sqrt{\frac{2}{\pi}}e^{-\frac{\lambda_{l,0}(\xi)^2}{2}} + \mathcal{O}(K^{-3/2})\right].$$

This yields (putting back the (w,ξ) dependency):

$$\prod_{l=1}^{K} J\left[\lambda_l(w,\xi),\hat{w},\hat{u}\right)] = e^{-\frac{1}{2}\hat{w}^2}\exp\left[-\frac{\hat{u}^2}{2} - i\hat{u}S_1 - i\sqrt{\frac{2}{\pi}}\hat{u}(\lambda_1 - i\hat{w})\Gamma_0 + \frac{1}{2}\hat{u}^2 S_2 + \mathcal{O}(K^{-1/2})\right], \tag{86}$$

in which we defined the following quantities, that only depend on ξ (recall (84))

$$w_\xi(\xi) \equiv \frac{1}{\sqrt{K}}\sum_{l=1}^{K}\xi_l, \qquad \Gamma_0(\xi) \equiv \frac{1}{K}\sum_{l=1}^{K} e^{-\frac{1}{2}\lambda_{l,0}(\xi)^2},$$

$$S_1(\xi) \equiv \frac{1}{\sqrt{K}}\sum_{l=1}^{K}\hat{H}(\lambda_{l,0}(\xi)), \qquad S_2(\xi) \equiv \frac{1}{K}\sum_{l=1}^{K}\hat{H}(\lambda_{l,0}(\xi))^2.$$

A detailed calculation actually shows that the previous expansion of (86) is valid up to $\mathcal{O}(K^{-1})$, and not only $\mathcal{O}(K^{-1/2})$. Recall also (81), in which one can now readily perform the integration over all variables $w,\hat{w},u,\hat{u}$ to obtain (dropping the ξ dependency in w_ξ,Γ_0,S_1,S_2):

$$I_C(y,\xi) = H\left[-y\frac{S_1 + \sqrt{\frac{2}{\pi}}w_\xi\Gamma_0\frac{\sqrt{q_d+q_a}-\sqrt{q_d}}{\sqrt{1-q_d}}}{\sqrt{1 - S_2 - \frac{2}{\pi}\Gamma_0^2\frac{q_a}{1-q_d}}}\right] + \mathcal{O}(K^{-1}), \tag{87}$$

in which $H(x) \equiv \int_x^\infty \mathcal{D}z = \frac{1}{2}\left[1 - \text{erf}(x/\sqrt{2})\right]$. Note that all quantities $w_\xi, \Gamma_0, S_1, S_2$ only depend on ξ via its empirical measure, which implies that the integration over $\xi \in \mathbb{R}^K$ will be tractable. We compute it in the following, using theoretical physics methods. We denote the quantity that appears in (87) as a function of $w_\xi, \Gamma_0, S_1, S_2$:

$$G(y, w_\xi, \Gamma_0, S_1, S_2) \equiv H\left[-y\frac{S_1 + \sqrt{\frac{2}{\pi}}w_\xi\Gamma_0\frac{\sqrt{q_d+q_a}-\sqrt{q_d}}{\sqrt{1-q_d}}}{\sqrt{1 - S_2 - \frac{2}{\pi}\Gamma_0^2\frac{q_a}{1-q_d}}}\right].$$

Introducing once again delta functions and their Fourier transforms for $w_\xi, \Gamma_0, S_1, S_2$, we write, starting from (87):

$$\begin{aligned} I_C &= \sum_{y=\pm 1}\int_{\mathbb{R}^K} \mathcal{D}\xi I_C(y,\xi)\log I_C(y,\xi) \\ &= \sum_{y=\pm 1}\int \frac{dw_\xi d\hat{w}_\xi}{2\pi}\frac{d\Gamma_0 d\hat{\Gamma}_0}{2\pi}\frac{dS_1 d\hat{S}_1}{2\pi}\frac{dS_2 d\hat{S}_2}{2\pi} e^{iw\hat{w}+i\Gamma_0\hat{\Gamma}_0+iS_1\hat{S}_1+iS_2\hat{S}_2}\, G(y, w_\xi, \Gamma_0, S_1, S_2) \\ &\qquad \times \log G(y, w_\xi, \Gamma_0, S_1, S_2)\left[\int_{\mathbb{R}^K}\mathcal{D}\xi e^{-i\hat{w}w_\xi(\xi)-i\hat{\Gamma}_0\Gamma_0(\xi)-i\hat{S}_1S_1(\xi)-i\hat{S}_2S_2(\xi)}\right] + \mathcal{O}(K^{-1}). \end{aligned} \tag{88}$$

The integral over ξ in (88) can be computed in the limit $K \to \infty$:

$$\begin{aligned} \Lambda &\equiv \int_{\mathbb{R}^K}\mathcal{D}\xi e^{-i\hat{w}w_\xi(\xi)-i\hat{\Gamma}_0\Gamma_0(\xi)-i\hat{S}_1S_1(\xi)-i\hat{S}_2S_2(\xi)} \\ &= \left[\int_{\mathbb{R}}\mathcal{D}\xi \exp\left[-i\frac{\hat{w}\xi}{\sqrt{K}} - i\frac{\hat{\Gamma}_0 e^{-\frac{q_d}{2(1-q_d)}\xi^2}}{K} - i\frac{\hat{S}_1\hat{H}\left[\sqrt{\frac{q_d}{1-q_d}}\xi\right]}{\sqrt{K}} - i\frac{\hat{S}_2\hat{H}\left[\sqrt{\frac{q_d}{1-q_d}}\xi\right]^2}{K}\right]\right]^K \end{aligned}$$

The large K expansion yields

$$\begin{aligned} \Lambda = \exp\Bigg\{ &-\frac{1}{2}\hat{w}^2 - i\hat{\Gamma}\sqrt{1-q_d} - \hat{S}_1\hat{w}\mathbb{E}\left[\xi\hat{H}\left(\sqrt{\frac{q_d}{1-q_d}}\xi\right)\right] \\ &- \left(\frac{1}{2}\hat{S}_1^2 + i\hat{S}_2\right)\mathbb{E}\left[\hat{H}\left(\sqrt{\frac{q_d}{1-q_d}}\xi\right)^2\right]\Bigg\} + \mathcal{O}(K^{-1}). \end{aligned}$$

The expectations are taken with respect to a real variable $\xi \sim \mathcal{N}(0,1)$. These expectations are known by properties of the error function:

$$\begin{aligned} \mathbb{E}\left[\hat{H}\left(\sqrt{\frac{q_d}{1-q_d}}\xi\right)^2\right] &= \frac{2}{\pi}\arcsin q_d\,, \\ \mathbb{E}\left[\xi\hat{H}\left(\sqrt{\frac{q_d}{1-q_d}}\xi\right)\right] &= \sqrt{\frac{2q_d}{\pi}}. \end{aligned}$$

One can now compute the integrals over the "hat" variables in (88). Denote $\Gamma_0^f \equiv \sqrt{\frac{2(1-q_d)}{\pi}}$, and $S_2^f \equiv$

$\frac{2}{\pi}\arcsin q_d$. This yields:

$$I_C = \int_{\mathbb{R}^2} \mathcal{D}w\mathcal{D}S_1\, G\left(y, w, \Gamma_0^f, \sqrt{\frac{2(\arcsin q_d - q_d)}{\pi}} S_1 + w\sqrt{\frac{2q_d}{\pi}}, S_2^f\right) \log G\left(y, w, \Gamma_0^f, \sqrt{\frac{2(\arcsin q_d - q_d)}{\pi}} S_1 + w\sqrt{\frac{2q_d}{\pi}}, S_2^f\right). \tag{89}$$

Note that

$$G\left(y, w, \Gamma_0^f, \sqrt{\frac{2(\arcsin q_d - q_d)}{\pi}} S_1 + w\sqrt{\frac{2q_d}{\pi}}, S_2^f\right) = H\left[-y\sqrt{\frac{2}{\pi}}\frac{\sqrt{\arcsin q_d - q_d}S_1 + w\sqrt{q_d + q_a}}{\sqrt{1 - \frac{2}{\pi}(q_a + \arcsin q_d)}}\right].$$

Making the change of variable $S_1^{new} = S_1 + w\frac{\sqrt{q_d+q_a}}{\sqrt{\arcsin q_d - q_d}}$ in (89), and defining $\gamma \equiv \frac{2}{\pi}(q_a + \arcsin q_d)$, one reaches:

$$I_C = \sum_{y=\pm 1} \int_{\mathbb{R}} \mathcal{D}x H\left[yx\sqrt{\frac{\gamma}{1-\gamma}}\right] \log H\left[yx\sqrt{\frac{\gamma}{1-\gamma}}\right] + \mathcal{O}(K^{-1}).$$

The two values of y contribute in the same way, which finally yields:

$$I_C = 2\int_{\mathbb{R}} \mathcal{D}x H\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right] \log H\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right] + \mathcal{O}(K^{-1}). \tag{90}$$

Note that the parameter γ is naturally bounded to the interval $[0, 1]$ by the conditions $0 \leq q_d \leq 1$ and $0 \leq q_a + q_d \leq 1$.

D.2 The Gaussian prior

The prior part I_P of the free entropy of (76) is very easy to evaluate in the Gaussian prior setting. We consider a prior with covariance matrix $Q_0 = I_K$ (we can simply rescale q by q/σ^2 in the final expression for a finite variance $Q_0 = \sigma^2 I_K$ as we already described). Performing the Gaussian integration in I_P in (76) yields:

$$I_P = \frac{K}{2}\hat{q}_d + \frac{1}{2}\hat{q}_a - \frac{K-1}{2}\log(1 + \hat{q}_d) - \frac{1}{2}\log\left(1 + \hat{q}_d + \hat{q}_a\right). \tag{91}$$

D.3 The fixed point equations

From the definition of the free entropy (76) and the expansions for I_P and I_C obtained in (90) and (91), one obtains the fixed point equations after having extremized over $\hat{q}_d$ and $\hat{q}_a$ (recall that $\alpha \equiv \lim \frac{m}{n}$):

$$\partial_{q_a}\left[I_G(q_d, q_a) + \alpha I_C(q_d, q_a)\right] = 0, \tag{92}$$

$$\partial_{q_d}\left[I_G(q_d, q_a) + \alpha I_C(q_d, q_a)\right] = 0, \tag{93}$$

with $I_G(q_d, q_a)$ defined as:

$$I_G(q_d, q_a) \equiv \frac{1}{2}\left[q_a + Kq_d\right] - \frac{K-1}{2}\log\left[\frac{1}{1-q_d}\right] - \frac{1}{2}\log\left[\frac{1}{1-q_a-q_d}\right],$$

$$I_C(q_d, q_a) = 2\int_{\mathbb{R}} \mathcal{D}x H\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right] \log H\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right],$$

and recall that $\gamma \equiv \frac{2}{\pi}(q_a + \arcsin q_d)$.

The fixed point equations (92), (93) have different behaviors depending on the scaling of α with the hidden layer size K. We detail these different behaviors in the following paragraphs.

D.3.1 Regime $\alpha = o_{K\to\infty}(K)$

In this regime (which in particular contains the case in which α stays of order 1 when $K \to \infty$), the fixed point equations (92), (93) can be simplified as:

$$\begin{cases} q_d = 0, \\ q_a = 2\alpha(1-q_a)\frac{\partial \mathcal{I}_C}{\partial q_a}. \end{cases} \tag{94}$$

D.3.2 Regime $\alpha = \Theta_{K\to\infty}(K)$

In this regime, we naturally define $\widetilde{\alpha}K \equiv \alpha/K$, such that $\widetilde{\alpha}$ will remain of order 1. One can show that the solutions of the fixed point equations (92), (93) must satisfy the following scaling : $q_a + q_d = 1 - \frac{\chi}{K}$, with $\chi \geq 0$ a reaching a finite value when $K \to \infty$. The fixed point equations in terms of χ and q_d read:

$$\begin{cases} q_d = 2(1-q_d)\left(\frac{1}{\sqrt{1-q_d^2}} - 1\right)\widetilde{\alpha}\frac{\partial \mathcal{I}_C}{\partial q_a}, \\ \chi^{-1} = 2\widetilde{\alpha}\frac{\partial \mathcal{I}_C}{\partial q_a}. \end{cases} \tag{95}$$

Note that the State Evolution (SE) computation of Figure 2 was performed by solving the fixed point equations (94) and (95) (depending on the regime of α).

The stability of the $q_d = 0$ solution: It is easy to show that (95) always admit what we call a *non-specialized solution*, i.e. a solution with $q_d = 0$. This solution stops to be optimal in term of the free energy at a finite $\widetilde{\alpha}_{\text{spec}} \simeq 7.65$. However, one can show that this solution will remain *linearly* stable for every $\widetilde{\alpha}$. Actually, it is linearly stable in the much broader regime $\alpha = o(K^2)$. Going back to the initial formulation of the fixed point equations (92),(93), and adding the correct time indices to iterate them, one obtains:

$$q_d^{t+1} = \frac{F(q_d^t, q_a^t)}{1 + F(q_d^t, q_a^t)}, \tag{96}$$

$$q_a^{t+1} = \frac{G(q_d^t, q_a^t)}{\left(1 + F(q_d^t, q_a^t)\right)\left(1 + F(q_d^t, q_a^t)G(q_d^t, q_a^t)\right)}, \tag{97}$$

with F and G defined as:

$$F(q_d, q_a) \equiv \frac{2\alpha}{K-1}\left[\partial_{q_d} I_C - \partial_{q_a} I_C\right], \tag{98}$$

$$G(q_d, q_a) \equiv \frac{2\alpha K}{K-1}\left[\partial_{q_a} I_C - \frac{1}{K}\partial_{q_d} I_C\right]. \tag{99}$$

We focus on the behavior of (96) around $q_d = 0$. Given our previous expansion of I_C in the $K \to \infty$ limit, and (98), one easily sees that for $\alpha = o_{K\to\infty}(K^2)$, $\frac{\partial F}{\partial q_d}|_{q_d=0} \to_{K\to\infty} 0$, which means the $q_d = 0$ solution always remains linearly stable.

However, assume now that $\alpha = \Theta(K^2)$. Performing a similar calculation to the one shown in sec. D.1,

one can show the following expansion:

$$I_C(q_d, q_a) = I_C^{(0)}(q_d, q_a) + \frac{1}{K} I_C^{(1)}(q_d, q_a) + \mathcal{O}\left(\frac{1}{K^2}\right).$$

The term of $\frac{\partial F}{\partial q_d}|_{q_d=0}$ arising from $I_C^{(1)}$ will thus have a possibly non-zero contribution in the $K \to \infty$ limit, as seen from (98).

To summarize, the non-specialized solution always remains linearly stable in the large K limit at least for $\alpha \ll K^2$. This implies that in this regime, Approximate Message Passing can not escape the non-specialized fixed point to find the specialized solution, as seen in Fig. 3. For α of order larger than K^2, one would have to explicitly compute $I_C^{(1)}$ in order to check that $\frac{\partial F}{\partial q_d}|_{q_d=0} \neq 0$ to show that the non-specialized solution is indeed linearly unstable. This tedious calculation is left for future work.

D.4 The generalization error at large K

Recall the definition of the generalization error in (78). From the remarks of section C, one can compute it at large K by applying the same techniques used to compute the channel integral I_C in sec. D.1. One obtains after a tedious, yet straightforward, calculation:

$$\epsilon_g^{\text{Bayes}} = \frac{1}{2}\epsilon_g^{\text{Gibbs}} = \frac{1}{\pi}\arccos\left[\frac{2}{\pi}\left(q_a + \arcsin q_d\right)\right] + \mathcal{O}(K^{-1}). \tag{100}$$

This expression is the one used in the computation of the generalization error in the left panel of Fig. 3.

E Linear networks show no specialization

An easy yet interesting case is a linear network with identical weights in the second layer and a final output function $\sigma : \mathbb{R} \to \mathbb{R}$, i.e a network in which $\varphi_{\text{out}}(\mathsf{h}) = \sigma\left(\frac{1}{\sqrt{K}}\sum_{l=1}^K h_l\right)$. For clarity, in this section, we decompose the channel as $P_{\text{out}}(y|\varphi_{\text{out}}(Z))$ for $Z \in \mathbb{R}^K$ instead of $P_{\text{out}}(y|Z)$. We will compute the channel integral I_C of the replica solution (76). For simplicity, we assume that $Q^0 = \mathbb{1}_K$ the identity matrix (i.e w has identity covariance matrix under P_0). Note that (76) gives I_C as $I_C = \int_{\mathbb{R}} dy \int_{\mathbb{R}^K} \mathcal{D}\xi I_C(y,\xi)\log I_C(y,\xi)$. One can easily derive:

$$I_C(y,\xi) = e^{-\frac{1}{2}\xi^\intercal(\mathbb{1}_K - q)^{-1}q\xi}\int_{\mathbb{R}^2}\frac{du d\hat{u}}{2\pi}e^{iu\hat{u}}P_{\text{out}}(y|\sigma(u)) \\ \times \int_{\mathbb{R}^K}\frac{dZ}{\sqrt{(2\pi)^K\det(\mathbb{1}_K - q)}}e^{-\frac{1}{2}Z^\intercal(\mathbb{1}_K-q)^{-1}Z + Z^\intercal X(\hat{u},xi)},$$

in which we denoted $X(\hat{u}, xi) \triangleq (\mathbb{1}_K - q)^{-1}q^{1/2}\xi - \frac{i\hat{u}}{\sqrt{K}}\mathsf{1}_K$, with the unit vector $\mathsf{1}_K = (1)_{l=1}^K$. The inner integration over Z can be done, as well as the integration over $\hat{u}$:

$$I_C(y,\xi) = \frac{1}{\sqrt{1 - \frac{1}{K}\mathsf{1}_K^\intercal q \mathsf{1}_K}}\int_{\mathbb{R}}\frac{du}{\sqrt{2\pi}}P_{\text{out}}(y|\sigma(u))\exp\left[-\frac{\left(u - \frac{1}{\sqrt{K}}\mathsf{1}_R^\intercal q^{1/2}\xi\right)^2}{2\left(1 - \frac{1}{K}\mathsf{1}_K^\intercal q \mathsf{1}_K\right)}\right].$$

So we can formally write the total dependency of $I_C(y,\xi)$ on ξ and on q as

$$I_C(y,\xi) = I_C\left(y, \frac{1}{\sqrt{K}}\mathsf{1}_K^\intercal q^{1/2}\xi, \frac{1}{K}\mathsf{1}_K^\intercal q \mathsf{1}_K\right).$$

Note that we have the following identity, for any fixed vector $x \in \mathbb{R}^K$ and smooth real function F:

$$\int_{\mathbb{R}^K} \mathcal{D}\xi F(x^\intercal \xi) = \frac{1}{\sqrt{2\pi x^\intercal x}} \int_{\mathbb{R}} du F(u) e^{-\frac{u^2}{2x^\intercal x}}. \tag{101}$$

In the end, if we denote $\Gamma(q) \triangleq \frac{1}{K} 1_K^\intercal q 1_K$, we have:

$$I_C = \int_{\mathbb{R}} dy \frac{1}{\sqrt{2\pi\Gamma(q)}} \int_{\mathbb{R}} dv e^{-\frac{v^2}{2\Gamma(q)}} I_C(v,y) \log I_C(v,y), \tag{102}$$

$$I_C(v,y) \equiv \frac{1}{\sqrt{2\pi(1-\Gamma(q))}} \int_{\mathbb{R}} du\, P_{\text{out}}(y|\sigma(u)) \exp\left[-\frac{1}{2\,(1-\Gamma(q))}\,(u-v)^2\right]. \tag{103}$$

Note that by hypothesis, both q and $\mathbb{1}_K - q$ are positive matrices, so $0 \leq \Gamma(q) \leq 1$. As these equations show, I_C only depends on $\Gamma(q) = K^{-1} \sum_{l,l'} q_{ll'}$. From this one easily sees that extremizing over q implies that the optimal $\hat{q}$ satisfies $\hat{q}_{ll'} = \hat{q}/K$ for some real $\hat{q}$. Subsequently, all $q_{ll'}$ are also equal to a single value, that we can denote $\frac{q}{K}$. This shows that this network never exhibits a specialized solution.

F Update functions and AMP derivation

AMP can be seen as Taylor expansion of the loopy belief-propagation (BP) approach [13, 14, 55], similar to the so-called Thouless-Anderson-Palmer equation in spin glass theory [35]. While the behaviour of AMP can be rigorously studied [17, 18, 56], it is useful and instructive to see how the derivation can be performed in the framework of belief-propagation and the cavity method, as was pioneered in [36, 38] for the single layer problem. The derivation uses the Generalized AMP notations of [16] and follows closely the one of [26].

F.1 De□nition of the update functions

Let's consider the distributions probabilities Q_{out} and Q_0, closely related to the inference problems eq. (3) and eq. (4):

$$Q_{\text{out}}(z;\omega,y,V) \equiv \frac{1}{\mathcal{Z}_{P_{\text{out}}}} e^{-\frac{1}{2}(z-\omega)^\intercal V^{-1}(z-\omega)} P_{\text{out}}(y|z); \qquad Q_0(W;\Sigma,T) \equiv \frac{1}{\mathcal{Z}_{P_0}} P_0(W) e^{-\frac{1}{2}W^\intercal \Sigma^{-1} W + T^\intercal \Sigma^{-1} W}.$$

We define the update functions g_{out}, $\partial_\omega g_{\text{out}}$, f_w and f_c, which will be useful later in the algorithm:

$$\begin{aligned}
g_{\text{out}}(\omega,y,V) &\equiv \partial_\omega \log(\mathcal{Z}_{P_{\text{out}}}) = V^{-1}\mathbb{E}_{Q_{\text{out}}}\left[z-\omega\right], \\
\partial_\omega g_{\text{out}}(\omega,y,V) &= V^{-1}\mathbb{E}_{Q_{\text{out}}}\left[(z-\omega)(z-\omega)^\intercal\right] - V^{-1} - g_{\text{out}} g_{\text{out}}^\intercal, \\
f_w(\Sigma,T) &\equiv \partial_{\Sigma^{-1}T} \log \mathcal{Z}_{P_0} = \mathbb{E}_{Q_0}[W], \\
f_c(\Sigma,T) &\equiv \partial_{\Sigma^{-1}T} f_w = \mathbb{E}_{Q_0}[WW^\intercal] - f_w f_w^\intercal.
\end{aligned}$$

Note that g_{out} is the mean of $V^{-1}(z-\omega)$ with respect tor Q_{out} and f_w the mean of Q_0.

F.2 Derivation of the Approximate Message Passing algorithm

F.2.1 Relaxed BP equations

Lets consider a set of messages $\{m_{i\to\mu}, \tilde{m}_{\mu\to i}\}_{i=1..n,\mu=1..m}$ on the bipartite factor graph corresponding to our problem Fig. 4. These messages correspond to the marginal probabilities of W_i if we remove the edges $i \to \mu$ or $\mu \to i$. The belief propagation (BP) equations (or sum-product equations) can be formulated as the following [14, 55], where $W_i = (w_{il})_{l=1..K} \in \mathbb{R}^K$:

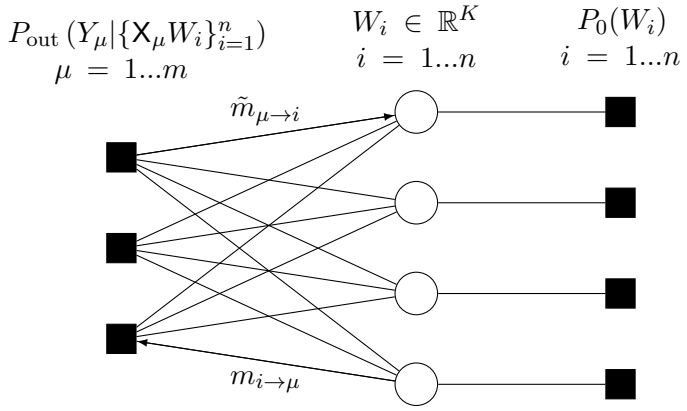

Figure 4: Factor graph representation of the committee machine (for $n = 4$ and $m = 3$). The variable (circle) $W_i \in \mathbb{R}^K$ needs to satisfy a prior constraint (square) P_0 and a constraint accounting for the fully connected layer, that correlates all the variables together.

$$\begin{cases} m_{i\to\mu}^{t+1}(W_i) = \dfrac{1}{\mathcal{Z}_{i\to\mu}} P_0(W_i) \displaystyle\prod_{k\neq\mu}^{m} \tilde{m}_{\nu\to i}^t(W_i)\,, \\ \tilde{m}_{\mu\to i}^t(W_i) = \dfrac{1}{\mathcal{Z}_{\mu\to i}} \displaystyle\int \prod_{j\neq i}^{n} dW_j P_{\text{out}}\left(Y_\mu|\frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_{\mu j}W_j\right) m_{j\to\mu}^t(W_j)\,. \end{cases} \tag{104}$$

The term inside P_{out} can be decouple using its K-dimensional Fourrier transform

$$P_{\text{out}}\left(Y_\mu|\frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_{\mu j}W_j\right) = \frac{1}{(2\pi)^{K/2}}\int_{\mathbb{R}^K} d\xi \exp\left(i\xi^\intercal\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n} X_{\mu j}W_j\right)\hat{P}_{\text{out}}(Y_\mu,\xi)\right)\,.$$

Injecting this representation in the BP equations, (104) becomes

$$\begin{aligned}\tilde{m}_{\mu\to i}^t(W_i) = \frac{1}{(2\pi)^{K/2}\mathcal{Z}_{\mu\to i}}\int_{\mathbb{R}^K} d\xi \hat{P}_{\text{out}}(Y_\mu,\xi)\exp\left(i\xi^\intercal\frac{1}{\sqrt{n}}X_{\mu i}W_i\right) \\ \times\prod_{j\neq i}^{n}\underbrace{\int_{\mathbb{R}^K} dW_j m_{j\to\mu}^t(W_j)\exp\left(i\xi^\intercal\frac{1}{\sqrt{n}}X_{\mu j}W_j)\right)}_{\equiv I_j}\,,\end{aligned}$$

and we define the mean and variance of the messages

$$\begin{cases} \hat{W}_{j\to\mu}^t \equiv \displaystyle\int_{\mathbb{R}^K} dW_j m_{j\to\mu}^t(W_j)W_j\,, \\ \hat{C}_{j\to\mu}^t \equiv \displaystyle\int_{\mathbb{R}^K} dW_j m_{j\to\mu}^t(W_j)W_jW_j^\intercal - \hat{W}_{j\to\mu}^t(\hat{W}_{j\to\mu}^t)^\intercal\,. \end{cases} \tag{105}$$

In the limit $n\to\infty$ the term I_j can be easily expanded and expressed using $\hat{W}$ and $\hat{C}$

$$I_j = \int_{\mathbb{R}^K} dW_j m_{j\to\mu}^t(W_j)\exp\left(i\xi^\intercal\frac{X_{\mu j}}{\sqrt{n}}W_j)\right) \simeq \exp\left(i\frac{X_{\mu j}}{\sqrt{n}}\xi^\intercal\hat{W}_{j\to\mu}^t - \frac{1}{2}\frac{X_{\mu j}^2}{n}\xi^\intercal\hat{C}_{j\to\mu}^t\,,\xi\right)\,,$$

and finally using the inverse Fourier transform, we obtain

$$\begin{aligned}
\tilde{m}^t_{\mu\to i}(W_i) &\simeq \frac{1}{(2\pi)^K \mathcal{Z}_{\mu\to i}} \int_{\mathbb{R}^K} dz P_{\rm out}(Y_\mu, z) \int_{\mathbb{R}^K} d\xi e^{-i\xi^\intercal z} e^{iX_{\mu i}\xi^\intercal W_i} \\
&\qquad \times \prod_{j\neq i}^n \exp\left(i\frac{X_{\mu j}}{\sqrt{n}} \xi^\intercal \hat{W}^t_{j\to\mu} - \frac{1}{2}\frac{X^2_{\mu j}}{n} \xi^\intercal \hat{C}^t_{j\to\mu}\xi \right) \\
&= \frac{1}{(2\pi)^K \mathcal{Z}_{\mu\to i}} \int_{\mathbb{R}^K} dz P_{\rm out}(Y_\mu, z) \int_{\mathbb{R}^K} d\xi e^{-i\xi^\intercal z} e^{iX_{\mu i}\xi^\intercal W_i} e^{i\xi^\intercal \sum_{j\neq i}^n \frac{X_{\mu j}}{\sqrt{n}} \hat{W}^t_{j\to\mu}} e^{-\frac{1}{2}\xi^\intercal \sum_{j\neq i}^n \frac{X^2_{\mu j}}{n} \hat{C}^t_{j\to\mu}\xi} \\
&= \frac{1}{(2\pi)^K \mathcal{Z}_{\mu\to i}} \int_{\mathbb{R}^K} dz P_{\rm out}(Y_\mu, z) \sqrt{\frac{(2\pi)^K}{\det(V^t_{i\mu})}} \underbrace{e^{-\frac{1}{2}\left(z - \frac{X_{\mu i}}{\sqrt{n}} W_i - \omega^t_{i\mu}\right)^\intercal (V^t_{i\mu})^{-1} \left(z - \frac{X_{\mu i}}{\sqrt{n}} W_i - \omega^t_{i\mu}\right)}}_{\equiv H_{i\mu}},
\end{aligned}$$

where we defined the mean and variance, depending on the node i

$$\omega^t_{i\mu} \equiv \frac{1}{\sqrt{n}} \sum_{j\neq i}^n X_{\mu j} \hat{W}^t_{j\to\mu}\,, \qquad V^t_{i\mu} \equiv \frac{1}{n} \sum_{j\neq i}^n X^2_{\mu j} \hat{C}^t_{j\to\mu}\,. \tag{106}$$

Again, in the limit $n \to \infty$, the term $H_{i\mu}$ can be expanded:

$$\begin{aligned}
H_{i\mu} \simeq e^{-\frac{1}{2}(z-\omega^t_{i\mu})^\intercal (V^t_{i\mu})^{-1}(z-\omega^t_{i\mu})} \Bigg(& 1 + \frac{X_{\mu i}}{\sqrt{n}} W_i^\intercal (V^t_{i\mu})^{-1}(z-\omega^t_{i\mu}) - \frac{1}{2}\frac{X^2_{\mu i}}{n} W_i^\intercal (V^t_{i\mu})^{-1} W_i \\
& + \frac{1}{2}\frac{X^2_{\mu i}}{n} W_i^\intercal (V^t_{i\mu})^{-1}(z-\omega^t_{i\mu})(z-\omega^t_{i\mu})^\intercal (V^t_{i\mu})^{-1} W_i \Bigg).
\end{aligned}$$

Gathering all pieces, the message $\tilde{m}_{\mu\to i}$ can be expressed using definitions of $g_{\rm out}$ and $\partial_\omega g_{\rm out}$

$$\begin{aligned}
\tilde{m}^t_{\mu\to i}(W_i) &\sim \frac{1}{\mathcal{Z}_{\mu\to i}} \Bigg\{ 1 + \frac{X_{\mu i}}{\sqrt{n}} W_i^\intercal g_{\rm out}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) + \frac{1}{2}\frac{X^2_{\mu i}}{n} W_i^\intercal g_{\rm out} g^\intercal_{\rm out}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) W_i + \\
&\quad \frac{1}{2}\frac{X^2_{\mu i}}{n} W_i^\intercal \partial_\omega g_{\rm out}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) W_i \Bigg\} \\
&= \frac{1}{\mathcal{Z}_{\mu\to i}} \left\{ 1 + W_i^\intercal B^t_{\mu\to i} + \frac{1}{2} W_i^\intercal B^t_{\mu\to i} (B^t_{\mu\to i})^\intercal (W_i) - \frac{1}{2} W_i^\intercal A^t_{\mu\to i} W_i \right\} \\
&= \sqrt{\frac{\det(A^t_{\mu\to i})}{(2\pi)^K}} \exp\left(-\frac{1}{2} \left(W_i^\intercal - (A^t_{\mu\to i})^{-1} B^t_{\mu\to i}\right)^\intercal A^t_{\mu\to i} \left(W_i^\intercal - (A^t_{\mu\to i})^{-1} B^t_{\mu\to i}\right) \right),
\end{aligned}$$

with the following definitions of $A_{\mu\to i}$ and $B_{\mu\to i}$:

$$B^t_{\mu\to i} \equiv \frac{X_{\mu i}}{\sqrt{n}} g_{\rm out}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}), \qquad A^t_{\mu\to i} \equiv -\frac{X^2_{\mu i}}{n} \partial_\omega g_{\rm out}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) \tag{107}$$

Using the set of BP equations (104), we can finaly close the set of equations only over $\{m_{i\to\mu}\}_{i\mu}$:

$$m^{t+1}_{i\to\mu}(W_i) = \frac{1}{\mathcal{Z}_{i\to\mu}} P_0(W_i) \prod_{\nu\neq\mu}^m \sqrt{\frac{\det(A^t_{\nu\to i})}{(2\pi)^K}} e^{-\frac{1}{2}\left(W_i - (A^t_{\nu\to i})^{-1} B^t_{\nu\to i}\right)^\intercal A^t_{\nu\to i} \left(W_i - (A^t_{\nu\to i})^{-1} B^t_{\nu\to i}\right)}.$$

In the end, computing the mean and variance of the product of gaussians, the messages are updated using

f_w and f_c:

$$\begin{cases} \hat{W}^{t+1}_{i\to\mu} = f_w(\Sigma^t_{\mu\to i}, T^t_{\mu\to i}) \,, \\ \hat{C}^{t+1}_{i\to\mu} = f_c(\Sigma^t_{\mu\to i}, T^t_{\mu\to i}) \,, \end{cases} \qquad \begin{cases} \Sigma^t_{\mu\to i} \equiv \left(\sum_{\nu\neq\mu}^{m} A^t_{\nu\to i} \right)^{-1} \,, \\ T^t_{\mu\to i} \equiv \Sigma^t_{\mu\to i} \left(\sum_{\nu\neq\mu}^{m} B^t_{\nu\to i} \right) \,. \end{cases} \tag{108}$$

Summary of the Relaxed BP set of equations:

In the end, using eq .(105,106,107, 108), relaxed BP equations can be written as the following set of equations:

$$\begin{cases} \omega^t_{i\mu} &= \sum_{j\neq i}^{n} \frac{X_{\mu j}}{\sqrt{n}} \hat{W}^t_{j\to\mu} \\ V^t_{i\mu} &= \sum_{j\neq i}^{n} \frac{X^2_{\mu j}}{n} \hat{C}^t_{j\to\mu} \\ B^t_{\mu\to i} &= \frac{X_{\mu i}}{\sqrt{n}} g_{\text{out}}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) \\ A^t_{\mu\to i} &= -\frac{X^2_{\mu i}}{n} \partial_\omega g_{\text{out}}(\omega^t_{i\mu}, Y_\mu, V^t_{i\mu}) \end{cases} \qquad \begin{cases} \Sigma^t_{\mu\to i} &= \left(\sum_{\nu\neq\mu}^{m} A^t_{\nu\to i} \right)^{-1} \\ T^t_{\mu\to i} &= \Sigma^t_{\mu\to i} \left(\sum_{\nu\neq\mu}^{m} B^t_{\nu\to i} \right) \\ \hat{W}^{t+1}_{i\to\mu} &= f_w(\Sigma^t_{\mu\to i}, T^t_{\mu\to i}) \\ \hat{C}^{t+1}_{i\to\mu} &= f_c(\Sigma^t_{\mu\to i}, T^t_{\mu\to i}) \end{cases} \tag{109}$$

F.2.2 Approximate Message Passing algorithm

The relaxed BP algorithm uses $\mathcal{O}(n^2)$ messages. However all the messages depend weakly on the target node. On a tree, the missing message is negligible, that allows us to expand the previous relaxed BP equations (109) to make appear the Onsager term at a previous time step, and reduce the number of messages to $\mathcal{O}(n)$. We define the following estimates and parameters based on the complete set of messages:

$$\begin{cases} \omega^t_\mu \equiv \sum_{j=1}^{n} \frac{X_{\mu j}}{\sqrt{n}} \hat{W}^t_{j\to\mu} \\ V^t_\mu \equiv \sum_{j=1}^{n} \frac{X^2_{\mu j}}{n} \hat{C}^t_{j\to\mu} \end{cases} \qquad \begin{cases} \Sigma^t_i \equiv \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} \right)^{-1} \\ T^t_i \equiv \Sigma^t_i \left(\sum_{\nu=1}^{m} B^t_{\nu\to i} \right) \end{cases} \tag{110}$$

Let's now expand the previous messages eq. (109), making appear these new target-independent messages:

- $\Sigma^t_{\mu\to i}$

$$\Sigma^t_{\mu\to i} = \left(\sum_{\nu\neq\mu}^{m} A^t_{\nu\to i} \right)^{-1} = \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} - A^t_{\mu\to i} \right)^{-1} = \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} \left(I_{K\times K} - \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} \right)^{-1} A^t_{\mu\to i} \right) \right)^{-1}$$
$$= \left(I_{K\times K} - \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} \right)^{-1} A^t_{\mu\to i} \right)^{-1} \left(\sum_{\nu=1}^{m} A^t_{\nu\to i} \right)^{-1} = \underbrace{\left(I_{K\times K} - \Sigma^t_i A^t_{\mu\to i} \right)^{-1}}_{\simeq I_{K\times K} + \Sigma^t_i A^t_{\mu\to i} + \mathcal{O}(n^{-1})} \Sigma^t_i \simeq \Sigma^t_i + \mathcal{O}\left(\frac{1}{n}\right)$$

- $T^t_{\mu\to i}$

$$T^t_{\mu\to i} = \Sigma^t_{\mu\to i} \left(\sum_{\nu\neq\mu}^{m} B^t_{\nu\to i} \right) = \left(\Sigma^t_i + \mathcal{O}\left(\frac{1}{n}\right) \right) \left(\sum_{\nu=1}^{m} B^t_{\nu\to i} - B^t_{\mu\to i} \right)$$
$$= T^t_i - \Sigma^t_i B^t_{\mu\to i} + \mathcal{O}\left(\frac{1}{n}\right)$$

- $\hat{W}_{i\to\mu}^{t+1}$

$$\begin{aligned}\hat{W}_{i\to\mu}^{t+1} &= f_w(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) = f_w\left(\Sigma_i^t, T_i^t - \Sigma_i^t B_{\mu\to i}^t\right) + \mathcal{O}\left(\frac{1}{n}\right)\\ &\simeq f_w\left(\Sigma_i^t, T_i^t\right) - \left.\frac{df_w}{dT}\right|_{\left(\Sigma_i^t, T_i^t\right)} \Sigma_i^t B_{\mu\to i}^t\\ &= \underbrace{f_w\left(\Sigma_i^t, T_i^t\right)}_{=\hat{W}_i^{t+1}} - \left(\Sigma_i^t\right)^{-1} \underbrace{f_c\left(\Sigma_i^t, T_i^t\right)\Sigma_i^t}_{=\hat{C}_i^{t+1}} \underbrace{B_{\mu\to i}^t}_{\simeq \frac{X_{\mu i}}{\sqrt{n}} g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t)}\\ &= \hat{W}_i^{t+1} - \frac{X_{\mu i}}{\sqrt{n}}\left(\Sigma_i^t\right)^{-1}\hat{C}_i^{t+1}\Sigma_i^t g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t) + \mathcal{O}\left(\frac{1}{n}\right)\end{aligned}$$

- $\hat{C}_{i\to\mu}^{t+1}$

  Let's denote for convenience, $\mathcal{E} = \left(\Sigma_i^t\right)^{-1}\hat{C}_i^{t+1}\Sigma_i^t g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t)$. Then

$$\begin{aligned}\hat{C}_{i\to\mu}^{t+1} &= \mathbb{E}_{Q_0}\left[\hat{W}_{i\to\mu}^t(\hat{W}_{i\to\mu}^t)^\intercal\right] - \mathbb{E}_{Q_0}\left[\hat{W}_{i\to\mu}^t\right]\mathbb{E}_{Q_0}\left[\hat{W}_{i\to\mu}^t\right]^\intercal\\ &= \mathbb{E}_{Q_0}\left[\left(\hat{W}_i^t - \frac{X_{\mu i}}{\sqrt{n}}\mathcal{E}\right)\left(\hat{W}_i^t - \frac{X_{\mu i}}{\sqrt{n}}\mathcal{E}\right)^\intercal\right] - \mathbb{E}_{Q_0}\left[\hat{W}_i^t - \frac{X_{\mu i}}{\sqrt{n}}\mathcal{E}\right]\mathbb{E}_{Q_0}\left[\hat{W}_i^t - \frac{X_{\mu i}}{\sqrt{n}}\mathcal{E}\right]^\intercal\\ &= \mathbb{E}_{Q_0}\left[\hat{W}_i^t(\hat{W}_i^t)^\intercal\right] - \mathbb{E}_{Q_0}\left[\hat{W}_i^t\right]\mathbb{E}_{Q_0}\left[\hat{W}_i^t\right]^\intercal + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) = \hat{C}_i^{t+1} + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

- $g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t)$

$$\begin{aligned}g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) &= g_{\text{out}}\left(\omega_\mu^t - \frac{X_{\mu i}}{\sqrt{n}}\hat{W}_{i\to\mu}^t, Y_\mu, V_\mu^t - \frac{X_{\mu i}^2}{n}\hat{C}_{i\to l}^t\right)\\ &= g_{\text{out}}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) - \frac{X_{\mu i}}{\sqrt{n}}\frac{\partial g_{\text{out}}}{\partial\omega}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) \underbrace{\hat{W}_{i\to\mu}^t}_{=\hat{W}_i^t + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)} + \mathcal{O}\left(\frac{1}{n}\right)\\ &= g_{\text{out}}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) - \frac{X_{\mu i}}{\sqrt{n}}\frac{\partial g_{\text{out}}}{\partial\omega}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right)\hat{W}_i^t + \mathcal{O}\left(\frac{1}{n}\right)\end{aligned}$$

- V_μ^t

$$V_\mu^t = \sum_{i=1}^n \frac{X_{\mu i}^2}{n}\hat{C}_{i\to l}^t = \sum_{i=1}^n \frac{X_{\mu i}^2}{n}\hat{C}_i^t + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)$$

- ω_μ^t

$$\begin{aligned}\omega_\mu^t &= \sum_{i=1}^n \frac{X_{\mu i}}{\sqrt{n}}\hat{W}_{i\to\mu}^t = \sum_{i=1}^n \frac{X_{\mu i}}{\sqrt{n}}\left(\hat{W}_i^t - X_{\mu i}\left(\Sigma_i^{t-1}\right)^{-1}\hat{C}_i^t\Sigma_i^{t-1}g_{\text{out}}(\omega_\mu^{t-1}, Y_\mu, V_\mu^{t-1}) + \mathcal{O}\left(\frac{1}{n}\right)\right)\\ &= \sum_{i=1}^n \frac{X_{\mu i}}{\sqrt{n}}\hat{W}_i^t - \sum_{i=1}^n \frac{X_{\mu i}^2}{n}\left(\Sigma_i^{t-1}\right)^{-1}\hat{C}_i^t\Sigma_i^{t-1}g_{\text{out}}(\omega_\mu^{t-1}, Y_\mu, V_\mu^{t-1}) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\end{aligned}$$

- $\left(\Sigma_i^t\right)^{-1}$

$$\left(\Sigma_i^t\right)^{-1} = \sum_{\mu=1}^{m} A_{\mu\to i}^t = -\sum_{\mu=1}^{m} X_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) = -\sum_{\mu=1}^{m} X_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)$$

- T_i^t

$$\begin{aligned} T_i^t &= \Sigma_i^t \left(\sum_{\mu=1}^{m} B_{\mu\to i}^t \right) = \Sigma_i^t \sum_{\mu=1}^{m} \frac{X_{\mu i}}{\sqrt{n}} g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) \\ &= \Sigma_i^t \sum_{\mu=1}^{m} \frac{X_{\mu i}}{\sqrt{n}} \left(g_{\text{out}}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) - \frac{X_{\mu i}}{\sqrt{n}} \frac{\partial g_{\text{out}}}{\partial \omega}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) \hat{W}_i^t + \mathcal{O}\left(\frac{1}{n}\right) \right) \\ &= \Sigma_i^t \left(\sum_{\mu=1}^{m} \frac{X_{\mu i}}{\sqrt{n}} g_{\text{out}}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) - \frac{X_{\mu i}^2}{n} \frac{\partial g_{\text{out}}}{\partial \omega}\left(\omega_\mu^t, Y_\mu, V_\mu^t\right) \hat{W}_i^t \right) + \mathcal{O}\left(\frac{1}{n^{3/2}}\right) \end{aligned}$$

The AMP algorithm follows naturally the rBP updates (109) using the expanded estimates of the mean and variance ω_μ, V_μ, T_i and Σ_i, and finally reads in pseudo language:

Algorithm 2 Approximate Message Passing for the committee machine

Input: vector $Y \in \mathbb{R}^m$ and matrix $X \in \mathbb{R}^{m\times n}$:
Initialize: $g_{\text{out},\mu} = 0, \Sigma_i = I_{K\times K}$ for $1 \le i \le n$ and $1 \le \mu \le m$ at $t = 0$.
Initialize: $\hat{W}_i \in \mathbb{R}^K$ and $\hat{C}_i, \partial_\omega g_{\text{out},\mu} \in \mathcal{S}_K^+$ for $1 \le i \le n$ and $1 \le \mu \le m$ at $t = 1$.
repeat
Update of the mean $\omega_\mu \in \mathbb{R}^K$ and covariance $V_\mu \in \mathcal{S}_K^+$:
$\omega_\mu^t = \sum_{i=1}^{n} \left(\frac{X_{\mu i}}{\sqrt{n}} \hat{W}_i^t - \frac{X_{\mu i}^2}{n} \left(\Sigma_i^{t-1}\right)^{-1} \hat{C}_i^t \Sigma_i^{t-1} g_{\text{out},\mu}^{t-1} \right) \quad | \quad V_\mu^t = \sum_{i=1}^{n} \frac{X_{\mu i}^2}{n} \hat{C}_i^t$
Update of $g_{\text{out},\mu} \in \mathbb{R}^K$ and $\partial_\omega g_{\text{out},\mu} \in \mathcal{S}_K^+$:
$g_{\text{out},\mu}^t = g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t) \quad | \quad \partial_\omega g_{\text{out},\mu}^t = \partial_\omega g_{\text{out}}(\omega_\mu^t, Y_\mu, V_\mu^t)$
Update of the mean $T_i \in \mathbb{R}^K$ and covariance $\Sigma_i \in \mathcal{S}_K^+$:
$T_i^t = \Sigma_i^t \Big(\sum_{\mu=1}^{m} \frac{X_{\mu i}}{\sqrt{n}} g_{\text{out},\mu}^t - \frac{X_{\mu i}^2}{n} \partial_\omega g_{\text{out},\mu}^t \hat{W}_i^t \Big) \quad | \quad \Sigma_i^t = -\Big(\sum_{\mu=1}^{m} \frac{X_{\mu i}^2}{n} \partial_\omega g_{\text{out},\mu}^t \Big)^{-1}$
Update of the estimated marginals $\hat{W}_i \in \mathbb{R}^K$ and $\hat{C}_i \in \mathcal{S}_K^+$:
$\hat{W}_i^{t+1} = f_w(\Sigma_i^t, T_i^t) \quad | \quad \hat{C}_i^{t+1} = f_c(\Sigma_i^t, T_i^t)$
$t = t + 1$
until Convergence on $\hat{W}, \hat{C}$.
Output: $\hat{W}$ and $\hat{C}$.

G State evolution equations from AMP

In this section, $W^\star$ denotes the ground truth weights of the teacher and we define the overlap parameters at time t, m^t, σ^t, q^t, Q and that respectively measure the correlation of the AMP estimator with the ground truth,

its variance and the norms of student and teacher weights:

$$\begin{cases} m^t \equiv \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{W}_i^t (W_i^\star)^\intercal \,, \\ q^t \equiv \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{W}_i^t (\hat{W}_i^t)^\intercal \,, \end{cases} \quad \text{and} \quad \begin{cases} \sigma^t \equiv \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{C}_i^t \,. \\ Q \equiv \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} W_i^\star (W_i^\star)^\intercal \,, \end{cases}$$

The aim is to derive the asymptotic behaviour of these overlap parameters, called state evolution. The idea is to compute the overlap distributions starting with the relaxed BP equations eq. (109).

G.1 Messages distribution

In order to get the asymptotic behaviour of the overlap parameters, we need first to compute the distribution of $\Sigma^t_{\mu\to i}$ and $T^t_{\mu\to i}$. Besides, we recall that in our model, the output has been generated by a teacher according to $Y_\mu = \varphi^0_{out}\left(\frac{1}{\sqrt{n}} W^\star X_\mu, A\right)$. We define $z_\mu \equiv \frac{1}{\sqrt{n}} W^\star X_\mu = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_{\mu i} W_i^\star$ and $z_{\mu\to i} \equiv \frac{1}{\sqrt{n}} \sum_{j\neq i}^n X_{\mu j} W_j^\star$. And it useful to recall $\mathbb{E}_X[X_{\mu i}] = 0$ and $\mathbb{E}_X[X^2_{\mu i}] = 1$.

- $\omega^t_{\mu\to i}$

Under belief propagation assumption messages are independent. $\omega^t_{\mu\to i}$ is thus the sum of independent variables and follows a gaussian distribution. Let's compute the first two moments, using expansions of the relaxed BP equations eq. (109):

$$\begin{aligned} \mathbb{E}_X\left[\omega^t_{\mu\to i}\right] &= \frac{1}{\sqrt{n}} \sum_{j\neq i}^{n} \mathbb{E}_X\left[X_{\mu j}\right] \hat{W}^t_{j\to\mu} = 0 \,, \\ \mathbb{E}_X\left[\omega^t_{\mu\to i}(\omega^t_{\mu\to i})^\intercal\right] &= \frac{1}{n} \sum_{j\neq i, k\neq i}^{n} \mathbb{E}_X\left[X_{\mu j} X_{\mu k}\right] \hat{W}^t_{j\to\mu} (\hat{W}^t_{k\to\mu})^\intercal = \sum_{j\neq i}^{n} \mathbb{E}_X\left[X^2_{\mu j}\right] \hat{W}_{j\to\mu} (\hat{W}_{j\to\mu})^\intercal \\ &= \frac{1}{n} \sum_{j\neq i}^{n} \hat{W}^t_{j\to\mu} (\hat{W}^t_{j\to\mu})^\intercal = \frac{1}{n} \sum_{i=1}^{n} \hat{W}^t_i (\hat{W}^t_i)^\intercal + \mathcal{O}\left(1/n^{3/2}\right) \underset{n\to\infty}{\longrightarrow} q^t \,. \end{aligned}$$

- z_μ

$$\begin{aligned} \mathbb{E}_X\left[z_\mu\right] &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}_X\left[X_{\mu i}\right] W_i^\star = 0 \,, \\ \mathbb{E}_{X,W^\star}\left[z_\mu z_\mu^\intercal\right] &= \mathbb{E}_{W^\star} \frac{1}{n} \sum_{j=1,k=1}^{n} \mathbb{E}_X\left[X_{\mu j} X_{\mu k}\right] W_j^\star (W_k^\star)^\intercal = \mathbb{E}_{W^\star} \frac{1}{n} \sum_{i=1}^{n} W_i^\star (W_i^\star)^\intercal \underset{n\to\infty}{\longrightarrow} Q \,. \end{aligned}$$

- z_μ and $\omega^t_{\mu\to i}$

$$\begin{aligned} \mathbb{E}_{X,W^\star}\left[\omega^t_{\mu\to i} z_\mu^\intercal\right] &= \mathbb{E}_{W^\star} \frac{1}{n} \sum_{j\neq i, k=1}^{n} \mathbb{E}_X\left[X_{\mu j} X_{\mu k}\right] \hat{W}^t_{j\to\mu} (W_k^\star)^\intercal = \mathbb{E}_{W^\star} \frac{1}{n} \sum_{j\neq i}^{n} \hat{W}^t_{j\to\mu} (W_j^\star)^\intercal \\ &= \mathbb{E}_{W^\star} \frac{1}{n} \sum_{i=1}^{n} \hat{W}^t_i (W_i^\star)^\intercal + \mathcal{O}\left(1/n^{3/2}\right) \underset{n\to\infty}{\longrightarrow} m^t \,. \end{aligned}$$

Hence asymptotically $(z_\mu, \omega^t_{\mu\to i})$ follow a Gaussian distribution with covariance matrix $\mathsf{Q}^t = \begin{bmatrix} Q & m^t \\ m^t & q^t \end{bmatrix}$.

• $V_{\mu\to i}$ concentrates around its mean:

$$\mathbb{E}_{X,W^\star}\left[V^t_{\mu\to i}\right] = \mathbb{E}_{W^\star}\frac{1}{n}\sum_{j\neq i}^{n}\mathbb{E}_X\left[X^2_{\mu j}\right]\hat{C}^t_{j\to\mu} = \mathbb{E}_{W^\star}\frac{1}{n}\sum_{j\neq i}^{n}\hat{C}^t_{j\to\mu} = \mathbb{E}_{W^\star}\frac{1}{n}\sum_{i}^{n}\hat{C}^t_{i} + \mathcal{O}\left(1/n^{3/2}\right) \underset{n\to\infty}{\longrightarrow} \sigma^t\,.$$

Let's define other order parameters, that will appear in the following:

$$\begin{cases} \hat{q}^t & \equiv \alpha\mathbb{E}_{\omega,z,A}\left[g_{out}(\omega,\varphi^0_{out}(z,A),\sigma^t)g_{out}(\omega,\varphi^0_{out}(z,A),\sigma^t)^\intercal\right]\,, \\ \hat{m}^t & \equiv \alpha\mathbb{E}_{\omega,z,A}\left[\partial_z g_{out}(\omega,\varphi^0_{out}(z,A),\sigma^t)\right]\,, \\ \hat{\chi}^t & \equiv \alpha\mathbb{E}_{\omega,z,A}\left[-\partial_\omega g_{out}(\omega,\varphi^0_{out}(z,A),\sigma^t)\right]\,. \end{cases}$$

• $T^t_{\mu\to i}$ can be expanded around $z_{\mu\to i}$:

$$\left(\Sigma^t_{\mu\to i}\right)^{-1}T^t_{\mu\to i} = \left(\sum_{\nu\neq\mu}^{m}B^t_{\nu\to i}\right) = \left(\sum_{\nu\neq\mu}^{m}\frac{1}{\sqrt{n}}X_{\nu i}g_{\rm out}(\omega^t_{\nu\to i},\varphi^0_{out}\left(\frac{1}{\sqrt{n}}\sum_{j\neq i}^{n}X_{\nu j}W^\star_j + X_{\nu i}W^\star_i, A\right), V^t_{\nu\to i})\right)$$

$$= \left(\sum_{\nu\neq\mu}^{m}\frac{1}{\sqrt{n}}X_{\nu i}g_{\rm out}(\omega^t_{\nu\to i},\varphi^0_{out}\left(z_{\nu\to i},A\right),V^t_{\nu\to i})\right) + \left(\sum_{\nu\neq\mu}^{m}\frac{1}{n}X^2_{\nu i}\partial_z g_{\rm out}(\omega^t_{\nu\to i},\varphi^0_{out}\left(z_{\nu\to i},A\right),V^t_{\nu\to i})\right)W^\star_i\,.$$

• $\Sigma^t_{\mu\to i}$

$$\begin{aligned}\left(\Sigma^t_{\mu\to i}\right)^{-1} &= \sum_{\nu\neq\mu}^{m}A^t_{\nu\to i} = -\sum_{\nu\neq\mu}^{m}\frac{1}{n}X^2_{\nu i}\partial_\omega g_{\rm out}(\omega^t_{\nu\to i},Y_\nu,V^t_{\nu\to i}) \\ &= -\sum_{\nu\neq\mu}^{m}\frac{1}{n}X^2_{\nu i}\partial_\omega g_{\rm out}(\omega^t_{\nu\to i},\varphi^0_{out}\left(z_{\nu\to i},A\right),V^t_{\nu\to i}) + \mathcal{O}\left(1/n^{3/2}\right)\,.\end{aligned}$$

Hence taking the average and the large size limit, the first moments of the variables $\Sigma^t_{\mu\to i}$ and $T^t_{\mu\to i}$ read:

$$\begin{cases} \mathbb{E}_{\omega,z,A,X}\left[\left(\Sigma^t_{\mu\to i}\right)^{-1}T^t_{\mu\to i}\right] \underset{n\to\infty}{\longrightarrow} \hat{m}^t W^\star_i\,, \\ \mathbb{E}_{\omega,z,A,X}\left[\left(\Sigma^t_{\mu\to i}\right)^{-1}T^t_{\mu\to i}\left(T^t_{\mu\to i}\right)^\intercal\left(\Sigma^t_{\mu\to i}\right)^{-1}\right] \underset{n\to\infty}{\longrightarrow} \hat{q}^t\,, \\ \mathbb{E}_{\omega,z,A,X}\left[\left(\Sigma^t_{\mu\to i}\right)^{-1}\right] \underset{n\to\infty}{\longrightarrow} \hat{\chi}^t\,. \end{cases}$$

And finally $T^t_{\mu\to i}\sim(\hat{\chi}^t)^{-1}\left(\hat{m}^t W^\star_i + (\hat{q}^t)^{1/2}\xi\right)$ with $\xi\sim\mathcal{N}(0,\mathbb{1})$ and $\left(\Sigma^t_{\mu\to i}\right)^{-1}\sim(\hat{\chi}^t)^{-1}$.

G.2 State evolution equations - Non Bayes optimal case

Let's define the following notations:

$$T^t[W^\star,\xi] \equiv (\hat{\chi}^t)^{-1}\left(\hat{m}^t W^\star + (\hat{q}^t)^{1/2}\xi\right)$$
$$\Sigma^t \equiv (\hat{\chi}^t)^{-1}$$

Gathering above results, the state evolution equations read:

$$\begin{cases} m^{t+1} = \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{W}_i^t (W_i^\star)^\intercal = \mathbb{E}_{W^\star,\xi} \left[f_w \left(\Sigma^t, T^t[W^\star, \xi] \right) (W^\star)^\intercal \right] \\ q^{t+1} = \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{W}_i^{t+1} (\hat{W}_i^{t+1})^\intercal = \mathbb{E}_{W^\star,\xi} \left[f_w \left(\Sigma^t, T^t[W^\star, \xi] \right) f_w \left(\Sigma^t, T^t[W^\star, \xi] \right)^\intercal \right] \\ \sigma^{t+1} = \mathbb{E}_{W^\star} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \hat{C}_i^{t+1} = \mathbb{E}_{W^\star,\xi} \left[f_c \left(\Sigma^t, T^t[W^\star, \xi] \right) \right] \end{cases}$$

and

$$\begin{cases} \hat{q}^t &= \alpha \mathbb{E}_{\omega,z,A} \left[g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t)^\intercal \right] \\ &= \alpha \int dP_A(A) \int dz d\omega \mathcal{N}\left(z, \omega; 0, \mathsf{Q}^t\right) g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t)^\intercal \\ \hat{m}^t &= \alpha \mathbb{E}_{\omega,z,A} \left[\partial_z g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) \right] \\ &= \alpha \int dP_A(A) \int dz d\omega \mathcal{N}\left(z, \omega; 0, \mathsf{Q}^t\right) \partial_z g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) \\ \hat{\chi}^t &= \alpha \mathbb{E}_{\omega,z,A} \left[-\partial_\omega g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) \right] \\ &= -\alpha \int dP_A(A) \int dz d\omega \mathcal{N}\left(z, \omega; 0, \mathsf{Q}^t\right) \partial_\omega g_{out}(\omega, \varphi_{out}^0(z,A), \sigma^t) \end{cases}$$

G.3 State evolution equations - Bayes optimal case

In the bayes optimal case, the student knows all the parameters of the teacher and then $P_0^\star = P_0$, $\varphi_{out}^0 = \varphi_{out}$, $m^t = q^t$ and $\hat{q}^t = \hat{m}^t = \hat{\chi}^t$, $\sigma^t = Q - q^t$ and then, naturally

$$T^t[W^\star, \xi] \equiv W^\star + (\hat{q}^t)^{-1/2} \xi \,,$$
$$\Sigma^t \equiv (\hat{q}^t)^{-1} \,.$$

In the Bayes-optimal case, the set of state evolution equations reduces and simplifies to:

$$\begin{cases} q^{t+1} = \mathbb{E}_{W^\star,\xi} \left[f_w \left(\Sigma^t, T^t[W^\star, \xi] \right) f_w \left(\Sigma^t, T^t[W^\star, \xi] \right)^\intercal \right] \,, \\ \hat{q}^t = \alpha \mathbb{E}_{\omega,z,A} \left[g_{out}(\omega, \varphi_{out}(z,A), \sigma^t) g_{out}(\omega, \varphi_{out}(z,A), \sigma^t)^\intercal \right] \,, \end{cases} \tag{111}$$

where $(z, \omega) \sim \mathcal{N}_{z,\omega}\left(0, 0; \mathsf{Q}^t\right)$ with $\mathsf{Q}^t = \begin{bmatrix} Q & q^t \\ q^t & q^t \end{bmatrix}$.

G.4 State evolution - Consistence between replicas and AMP - Bayes optimal case

State evolution - AMP

Using the change of variable $\xi \leftarrow \xi + \left(\hat{q}^t\right)^{1/2} W^\star$, eq. (111) becomes:

$$q^{t+1} = \mathbb{E}_\xi \left[\mathcal{Z}_{P_0} \left((\hat{q}^t)^{1/2} \xi, (\hat{q}^t)^{-1} \right) f_w \left((\hat{q}^t)^{1/2} \xi, (\hat{q}^t)^{-1} \right) f_w \left((\hat{q}^t)^{1/2} \xi, (\hat{q}^t)^{-1} \right)^\intercal \right]$$

In addition in the Bayes-optimal case, as:

$$\begin{cases} \mathbb{E}_X\left[\omega^t_{\mu\to i}(z_\mu - \omega^t_{\mu\to i})^\intercal\right] = m^t - q^t = 0 \\ \mathbb{E}_X[\omega^t_{\mu\to i}(\omega^t_{\mu\to i})^\intercal] = q^t \\ \mathbb{E}_X\left[(z^\intercal_\mu - \omega^t_{\mu\to i})(z_\mu - \omega^t_{\mu\to i})^\intercal\right] = Q - q^t \,, \end{cases}$$

the multivariate distribution can be written as a product: $\mathcal{N}_{z,\omega}\left(0,0;\mathsf{Q}^t\right) = \mathcal{N}_\omega\left(0,q^t\right)\mathcal{N}_z\left(\omega, Q-q^t\right)$. Hence, using $P_{\text{out}}(y|z) = \int dP_A(A)\delta\left(y - \varphi^0_{\text{out}}(z,A)\right)$, eq. (111) becomes:

$$\begin{aligned} \hat{q}^t &= \alpha\mathbb{E}_{\omega,z,A}\left[g_{out}(\omega,\varphi^0_{out}(z,A),Q-q^t)g_{out}(\omega,\varphi^0_{\text{out}}(z,A),Q-q^t)^\intercal\right] \\ &= \alpha\int dy\int d\omega\frac{e^{-\frac{1}{2}\omega^\intercal(q^t)^{-1}\omega}}{(2\pi)^{K/2}\det(q^t)^{1/2}}\int dz P_{\text{out}}(y|z)\frac{e^{-\frac{1}{2}(z-\omega)^\intercal(Q-q^t)^{-1}(z-\omega)}}{(2\pi)^{K/2}\det(Q-q^t)^{1/2}}g_{\text{out}}(\omega,y,Q-q^t)g_{\text{out}}(\omega,y,Q-q^t)^\intercal \\ &= \alpha\int dy\int D\xi\int dz P_{out}(y|z)\frac{e^{-\frac{1}{2}(z-\omega)^\intercal(Q-q^t)^{-1}(z-\omega)}}{(2\pi)^{K/2}\det(Q-q^t)^{1/2}}g_{\text{out}}((q^t)^{1/2}\xi,y,Q-q^t)g_{\text{out}}((q^t)^{1/2}\xi,y,Q-q^t)^\intercal \\ &= \alpha\mathbb{E}_{y,\xi}\left[\mathcal{Z}_{P_{\text{out}}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)g_{\text{out}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)g_{\text{out}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)^\intercal\right] \end{aligned}$$

Finally to summarize the state evolution equations can be written as:

$$\begin{cases} q^{t+1} = \mathbb{E}_\xi\left[\mathcal{Z}_{P_0}\left((\hat{q}^t)^{1/2}\xi,(\hat{q}^t)^{-1}\right)f_w\left((\hat{q}^t)^{1/2}\xi,(\hat{q}^t)^{-1}\right)f_w\left((\hat{q}^t)^{1/2}\xi,(\hat{q}^t)^{-1}\right)^\intercal\right] \\ \hat{q}^t = \alpha\mathbb{E}_{y,\xi}\left[\mathcal{Z}_{P_{\text{out}}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)g_{\text{out}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)g_{\text{out}}\left((q^t)^{1/2}\xi,y,Q-q^t\right)^\intercal\right] \end{cases} \tag{112}$$

State evolution - Replicas

Recall from sec. B, the free entropy eq. (76) reads

$$\begin{cases} \lim_{n\to\infty} f_n &= \text{extr}_{q,\hat{q}}\left\{-\frac{1}{2}\text{Tr}[q\hat{q}] + I_P + \alpha I_C\right\} \,, \\ I_P &\equiv \mathbb{E}_\xi\left[\mathcal{Z}_{P_0}(\hat{q}^{1/2}\xi,\hat{q}^{-1})\log(\mathcal{Z}_{P_0}(\hat{q}^{1/2}\xi,\hat{q}^{-1}))\right] \,, \\ I_C &\equiv \mathbb{E}_{\xi,y}\left[\mathcal{Z}_{P_{\text{out}}}(q^{1/2}\xi,y,Q-q)\log(\mathcal{Z}_{P_{\text{out}}}(q^{1/2}\xi,y,Q-q))\right] \,. \end{cases}$$

Taking the derivatives with respect to q and $\hat{q}$, using an integration by part and the following identities:

$$\begin{cases} \frac{\partial\mathcal{Z}_{P_{\text{out}}}}{\partial q} = -\frac{1}{2}q^{-1}e^{\frac{1}{2}\xi^\intercal\xi}\partial_\xi\left[e^{-\frac{1}{2}\xi^\intercal\xi}\partial_\xi\mathcal{Z}_{P_{\text{out}}}\right] \,, \\ \frac{\partial\mathcal{Z}_{P_0}}{\partial\hat{q}} = -\frac{1}{2}\hat{q}^{-1}e^{\frac{1}{2}\xi^\intercal\xi}\partial_\xi\left[e^{-\frac{1}{2}\xi^\intercal\xi}\partial_\xi\mathcal{Z}_{P_0}\right] \,, \end{cases}$$

the state evolution equations read:

$$\begin{cases} q = 2\frac{\partial I_P}{\partial\hat{q}} \\ \hat{q} = 2\alpha\frac{\partial I_C}{\partial q} \end{cases} \quad \text{with} \quad \begin{cases} \frac{\partial I_P}{\partial\hat{q}} = \frac{1}{2}\mathbb{E}_\xi\left[\mathcal{Z}_{P_0}(\hat{q}^{1/2}\xi,\hat{q}^{-1})f_w(\hat{q}^{1/2}\xi,\hat{q})f_w(\hat{q}^{1/2}\xi,\hat{q}^{-1})^\intercal\right] \\ \frac{\partial I_C}{\partial q} = \frac{1}{2}\mathbb{E}_{y,\xi}\left[\mathcal{Z}_{P_{\text{out}}}(q^{1/2}\xi,y,Q-q)g_{\text{out}}(q^{1/2}\xi,y,Q-q)g_{\text{out}}(q^{1/2}\xi,y,Q-q)^\intercal\right] \end{cases}$$

that simplifies and allows to recover the state evolutions equations directly derived from AMP eq. (112), but

without time indices

$$\begin{cases} q = \mathbb{E}_\xi \left[\mathcal{Z}_{P_0}(\hat{q}^{1/2}\xi, \hat{q}^{-1}) f_w(\hat{q}^{1/2}\xi, \hat{q}) f_w(\hat{q}^{1/2}\xi, \hat{q}^{-1})^\intercal \right] , \\ \hat{q} = \alpha \mathbb{E}_{y,\xi} \left[\mathcal{Z}_{P_{\text{out}}}(q^{1/2}\xi, y, Q-q) g_{\text{out}}(q^{1/2}\xi, y, Q-q) g_{\text{out}}(q^{1/2}\xi, y, Q-q)^\intercal \right] . \end{cases}$$

H Parity machine for $K = 2$

Although we mainly focused on the committee machine, another classical two-layers neural network is the parity machine [7] and our proof applies to this case as well. While learning is known to be computationally hard for general K, the case $K = 2$ is special, and in fact can be reformulated as a committee machine, where the sign activation function has been replaced by $\varphi_1(z) = \mathbb{1}(z \neq 0) - \mathbb{1}(z = 0)$:

$$Y_\mu = \text{sign}\Big[\prod_{l=1}^{K} \text{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W_{il}^* \Big) \Big] = \varphi_1 \Big[\sum_{l=1}^{K} \text{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W_{il}^* \Big) \Big] . \tag{113}$$

We have repeated our analysis for the $K = 2$ parity machine and the phase diagram is summarized in Fig. 5 where we show the generalization error and the elements of the overlap matrix for Gaussian (left) and binary weights (right), with the results of the AMP algorithm (points).

Below the specialization phase transition $\alpha < \alpha_{\text{spec}}$, the symmetry of the output imposes the non-specialized fixed point $q_{00} = q_{01} = 0$ to be the only solution, with $\alpha_{\text{spec}}^G(K = 2) \simeq 2.48$ and $\alpha_{\text{spec}}^B(K = 2) \simeq 2.49$. Above the specialization transition α_{spec}, the overlap becomes specialized with a non-trivial diagonal term.

Additionally, in the binary case, an information theoretical transition towards a perfect learning occurs at $\alpha_{\text{IT}}^B(K = 2) \simeq 2.00$, meaning that the perfect generalization fixed point ($q_{00} = 1, q_{01} = 0$) becomes the global optimizer of the free entropy. It leads to a first order phase transition of the AMP algorithm which retrieves the perfect generalization phase only at $\alpha_{\text{perf}}^B(K = 2) \simeq 3.03$. This is similar to what happens in single layer neural networks for the symmetric door activation function, see [11]. Again, these results for the parity machine emphasize a gap between information-theoretical and computational performance.

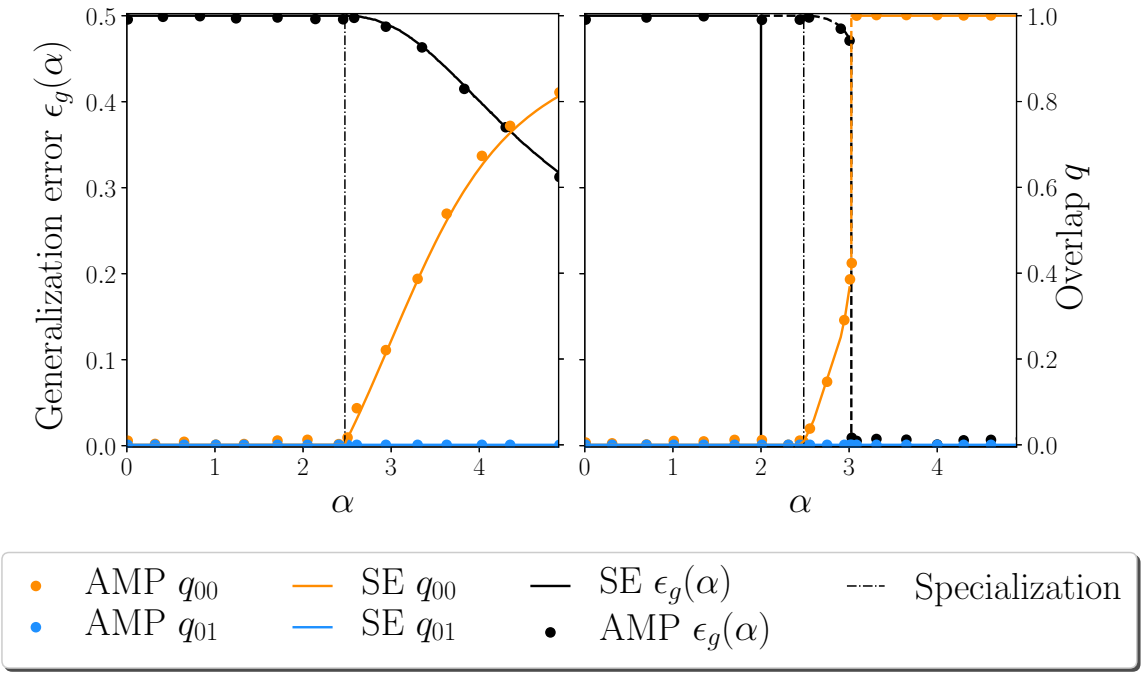

Figure 5: Similar plot as in Fig. 2 but for the parity machine with two hidden neurons. Value of the order parameter and the optimal generalization error for a parity machine with two hidden neurons with Gaussian weights (left) and binary/Rademacher weights (right). SE and AMP overlaps are respectively represented in full line and points.