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

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, Lenka Zdeborová

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. 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

† Laboratoire de Physique Statistique, CNRS & Sorbonnes Universités & École Normale Supérieure, PSL University, Paris, France.
⊗ Laboratoire de Théorie des Communications, Faculté Informatique et Communications, Ecole Polytechnique Fédérale de Lausanne, Suisse.
◊ Probability and Applications Group, School of Mathematical Sciences, Queen Mary University of London, United-Kingdom.
⋆ Institut de Physique Théorique, CNRS & CEA & Université Paris-Saclay, Saclay, France.

1 Introduction

While the traditional approach to learning and generalization follows the Vapnik-Cervenokis [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 $\alpha = m/n = \Theta(1)$ (i.e. scales as an order 1 constant) 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[1].

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.

[1]Note that many of these works study the "tree" committee machine, sometimes called committee machine with non-overlapping fields; we *do not* study this version here. We chose the version that is more closely related to currently used architectures.

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. 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. Gaussian coordinates $X_{\mu i} \sim \mathcal{N}(0,1)$, then she 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 $(Y_\mu, X_\mu)_{\mu=1}^{m}$ and she 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 $(Y_\mu, X_\mu)_{\mu=1}^{m}$ 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 $\alpha = m/n = \Theta(1)$, $K = \Theta(1)$ 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), 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 $\alpha > \alpha_{\text{spec}}$, 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 firm rigorous basis. Our second contribution is the design of AMP-type of algorithm that is able to achieve the optimal learning 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 larger 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 provide only a poor generalization that does not decay when increasing α. This strongly suggests that no efficient algorithm exists in this hard region and therefore there is a computational gap in learning in 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.

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 $\varphi_{\text{out}} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ 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. $X_{\mu i} \sim \mathcal{N}(0,1)$, and i.i.d. weights $W^*_{il} \sim P_0$. The student then learns W^* from the data $(Y_\mu, X_\mu)_{\mu=1}^m$ by computing marginal means of the posterior probability distribution (9).

Different scenarii fit into this general framework. Among those, the committee machine is obtained when choosing $\varphi_{\rm out}(h) = \text{sign}(\sum_{l=1}^K \text{sign}(h_l))$. Another model is given by the parity machine, when $\varphi_{\rm out}(h) = \prod_{l=1}^K \text{sign}(h_l)$, see e.g. [7], and we discuss this example further in appendix H. 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 [29, 30] or multi-vector compressed sensing [31].

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 and compact 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 $S_K^+(N) \equiv \{M \in S_K^+ \text{ s.t. } N - M \in \mathcal{S}_K^+\}$.

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

• **Input Gaussian channel:** 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

$$Y_0 = R^{1/2}W_0 + Z_0 \tag{3}$$

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

$$P(w|Y_0) = \frac{1}{\mathcal{Z}_{P_0}} P_0(w) e^{Y_0^\intercal R^{1/2} w - \frac{1}{2} w^\intercal R w}\,, \tag{4}$$

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

$$\psi_{P_0}(R) \equiv \mathbb{E}\ln \mathcal{Z}_{P_0} = \mathbb{E}\ln \int_{\mathbb{R}^K} dP_0(w) e^{Y_0^\intercal R^{1/2} w - \frac{1}{2} w^\intercal R w} \tag{5}$$

and involves K dimensional integrals.

• **Output channel:** 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

$$\widetilde{Y}_0 \sim P_{\rm out}(\cdot|q^{1/2}V + (\rho - q)^{1/2}U^*) \tag{6}$$

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

in $S_K^+(\rho)$. The associated posterior is

$$P(u|\widetilde{Y}_0, V) = \frac{1}{\mathcal{Z}_{P_{\rm out}}} \frac{e^{-\frac{1}{2}u^\intercal u}}{(2\pi)^{K/2}} P_{\rm out}\big(\widetilde{Y}_0|q^{1/2}V + (\rho-q)^{1/2}u\big)\,, \tag{7}$$

and the free entropy reads this time

$$\Psi_{P_{\rm out}}(q;\rho) \equiv \mathbb{E}\ln \mathcal{Z}_{P_{\rm out}} = \mathbb{E}\ln \int_{\mathbb{R}^K} \big(\prod_{i=1}^K du_i\big) \frac{e^{-\frac{1}{2}u^\intercal u}}{(2\pi)^{K/2}} P_{\rm out}\big(\widetilde{Y}_0|q^{1/2}V + (\rho-q)^{1/2}u\big) \tag{8}$$

(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:

$$P(\{w_{il}\}_{i,l=1}^{n,K} \mid \{Y_\mu, X_{\mu i}\}_{\mu,i=1}^{m,n}) = \frac{1}{\mathcal{Z}_n} \prod_{i=1}^n P_0(\{w_{il}\}_{l=1}^K) \prod_{\mu=1}^m P_{\rm out}\Big(Y_\mu\Big|\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}\Big\}_{l=1}^K\Big)\,, \tag{9}$$

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[2] free entropy is by definition

$$f_n \equiv \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_n = \frac{1}{n}\mathbb{E}\ln \int \prod_{i=1}^n dP_0(\{w_{il}\}_{l=1}^K) \prod_{\mu=1}^m P_{\rm out}\Big(Y_\mu\Big|\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{\mu i}w_{il}\Big\}_{l=1}^K\Big)\,. \tag{10}$$

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 discuss in the supplementary material (appendices sec. C) 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 rigourous proof of the corresponding free entropy formula using an interpolation method [32, 33, 24].

In order to formulate our rigorous results, we add a (arbitrarily small) Gaussian term $Z_\mu\sqrt{\Delta}$ to the first expression of the model (2), where $\Delta > 0$, $Z_\mu \sim \mathcal{N}(0,1)$, so that the channel kernel is ($u \in \mathbb{R}^K$)

$$P_{\rm out}(y|u) = \frac{1}{\sqrt{2\pi\Delta}} \int_{\mathbb{R}} dP_A(a) e^{-\frac{1}{2\Delta}(y-\varphi(u,a))^2}. \tag{11}$$

Theorem 3.1 (Replica formula). *Suppose (H1): The prior P_0 has bounded support in $\mathbb{R}^K$; (H2): The activation $\varphi_{\rm out} : \mathbb{R}^K \times \mathbb{R} \to \mathbb{R}$ is a bounded $\mathcal{C}^2$ function with bounded first and second derivatives w.r.t. its first argument (in $\mathbb{R}^K$-space); and (H3): For all $\mu = 1,\dots,m$ and $i = 1,\dots,n$ we have $X_{\mu i} \sim \mathcal{N}(0,1)$. Then for the model (2) with kernel (11), the $m, n \to \infty$ limit of the free entropy in the regime $\alpha = m/n = \Theta(1)$, $K = \Theta(1)$ is:*

$$\lim_{n\to\infty} f_n \equiv \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_n = \sup_{R\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} \Big\{\psi_{P_0}(R) + \alpha\Psi_{P_{\rm out}}(q;\rho) - \frac{1}{2}{\rm Tr}(Rq)\Big\}\,, \tag{12}$$

where $\Psi_{P_{\rm out}}(q;\rho)$ and $\psi_{P_0}(R)$ are the free entropies of simpler K-dimensional estimation problems (4) and (7).

This theorem extends the recent progress for generalized linear models of [11], which includes the case

[2] 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.

$K = 1$ 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 and the details can be found in the appendices sec. A. 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 $\mathrm{sign}(\cdot)$ activation) and (H3) extended to weight matrices with i.i.d. entries of zero mean, unit variance and finite third moment.

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 (12), we expect from the replica method that in the limit $n \to \infty, m/n \to \alpha$, with high probablity $\hat{W}^\intercal W^*/n \to q^*$. We refer to proposition 5.2 and to the proof in appendices 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

$$\frac{1}{2}\mathbb{E}\left[\left(\mathbb{E}_{W|Y,X}\left[\varphi_{\text{out}}(X_{\text{new}}W)|Y,X\right] - \varphi_{\text{out}}(X_{\text{new}}W^*)\right)^2\right] \xrightarrow[n\to\infty]{} \epsilon_g(q^*) \tag{13}$$

where W is distributed according to the posterior measure (9) (note that this Bayes-optimal computation differs from the so-called Gibbs estimator by a factor 2, see appendix sec. D). 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 $W^\intercal W^*$, which converges to q^*, and a direct algebraic computation gives a lengthy but explicit formula for $\epsilon_g(q^*)$, as shown in the appendices.

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 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 (as shown in appendices G.2) and also originate in earlier statistical physics works [34, 35, 36, 37, 38, 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.

The AMP algorithm is summarized by its pseudo-code in Algorithm 1, where the update functions g_{out}, $\partial_\omega g_{\text{out}}$, f_W and f_C are related, again, to the two auxiliary problems (4) and (7). The functions $f_W(\Sigma, T)$ and $f_C(\Sigma, T)$ are the mean and variance under the measure of the posterior (4) when $R = \Sigma^{-1}$ and $Y_0 = \Sigma^{1/2}T$, while $g_{\text{out}}(\omega_\mu, Y_\mu, V_\mu)$ is given by the expected mean of $V^{-1/2}u$ under the posterior (7) using $\widetilde{Y}_0 = Y_\mu$, $\rho - q = V_\mu$ and $q^{1/2}V = \omega_\mu$ (see appendix G.1 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

$$Y^t_{\text{new}} = \int \mathrm{d}y \Big(\prod_{l=1}^{K} \mathrm{d}z_l\Big)\, y P_{\text{out}}(y|\{z_l\}_{l=1}^K)\mathcal{N}(z; \omega^t_{\text{new}}, V^t_{\text{new}})\,, \tag{14}$$

where $\omega^t_{\text{new}} = \sum_{i=1}^n X_{\text{new},i}\hat{W}^t_i$ is the mean of the normally distributed variable $z \in \mathbb{R}^K$, and $V^t_{\text{new}} = \rho - q^t_{\text{AMP}}$

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: $\hat{W}_i$, $g_{\text{out},\mu} \in \mathbb{R}^K$ 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 $t = 0$.
repeat
Update of the mean $\omega_\mu \in \mathbb{R}^K$ and covariance $V_\mu \in \mathcal{S}_K^+$:

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

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

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

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

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

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

$$\hat{W}_i^{t+1} = f_W(\Sigma_i^t, T_i^t) \qquad \hat{C}_i^{t+1} = f_C(\Sigma_i^t, T_i^t)$$

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

is the $K \times K$ covariance matrix (see below for the definition of q_{AMP}^t). We provide a demo of the algorithm on github [39].

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 the AMP estimate $\hat{W}_t$, defined as $q_{\text{AMP}}^t \equiv (\hat{W}^t)^\intercal W^*/n$ via:

$$q_{\text{AMP}}^{t+1} = 2\frac{\partial \psi_{P_0}}{\partial R}(R_{\text{AMP}}^t)\,, \qquad R_{\text{AMP}}^{t+1} = 2\alpha \frac{\partial \Psi_{P_{\text{out}}}}{\partial q}(q_{\text{AMP}}^t; \rho)\,. \tag{15}$$

The fixed points of these equations correspond to the critical points of the replica free entropy (12).

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, i.e. when model (1) is simply

$$Y_\mu = \text{ sign}\Big[\text{ sign}\Big(\sum_{i=1}^n X_{\mu i} W_{i1}^*\Big) + \text{ sign}\Big(\sum_{i=1}^n X_{\mu i} W_{i2}^*\Big)\Big], \tag{16}$$

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. 1 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 center panel they are binary/Rademacher (recall that (H3) in Theorem 3.1 can be relaxed to include this case, see [11]). The full line is obtained from the fixed point of the state evolution (SE) of the AMP algorithm (15), corresponding to the

extremizer of the replica free entropy (12). The points are results of the AMP algorithm run till convergence averaged over 10 instances of size $n = 10^4$. As expected we observe excellent agreement between the SE and AMP.

In both left and center panels of Fig. 1 we observe the so-called *specialization* phase transition. Indeed (15) has two types of fixed points: A *non-specialized* fixed point where every 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 non-linearly separable model. The specialized fixed point is associated with lower error that the non-specialized one (as one can see in Fig. 1). 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. 1 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 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. 1). For completeness we remind that in the case of $K = 1$ corresponding to single-layer neural network no such specialization transition exists. We show (see appendices sec. F) 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 center part of Fig. 1 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 optimizer of the replica free entropy (not the local optimizer 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. 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

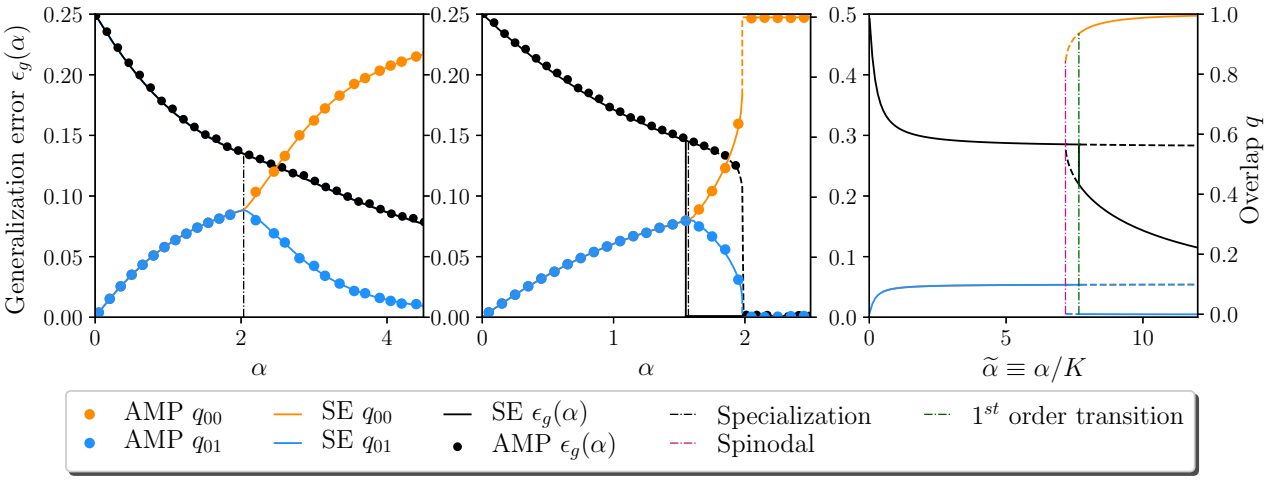

Figure 1: Value of the order parameter and the optimal generalization error for a committee machine with two hidden neurons with Gaussian weights (left), binary/Rademacher weights (center), and for Gaussian weights in the large number of hidden units limit (right). These are shown as a function of the ratio $\alpha = m/n$ between the number of samples m and the dimensionality n. Lines are obtained from the state evolution equations (dominating solution is shown in full line), data-points from the AMP algorithm (see implementation and demo on github [39]) averaged over 10 instances of the problem of size $n = 10^4$. q_{00} and q_{01} denote respectively diagonal and off-diagonal overlaps, and their value is to be read on the labels on the far-right of the figure.

positive, this scaling implies (see appendices sec. E) that $0 \leq q_d \leq 1$ and $0 \leq q_a + q_d \leq 1$, as it should. We also detail the corresponding expansion 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 plot of Fig. 1 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. As discussed in [19] it can be written in a closed form as $\epsilon_g = \pi^{-1} \arccos\left[2\left(q_a + \arcsin q_d\right)/\pi\right]$. The specialization transition arises for $\alpha = \Theta(K)$ so we define $\widetilde{\alpha} \equiv \alpha/K$. The specialization is now a first 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 $1.25/\widetilde{\alpha}$.

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 first 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 in sec. E.1 of the appendices, 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. 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 Sketch of proof of Theorem 3.1

In order to avoid confusions we denote K-dimensional column vectors by underlined letters. In particular we set $\underline{W}_i^* = (W_{il}^*)_{l=1}^K$, $\underline{w}_i^* = (w_{il}^*)_{l=1}^K$. 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 $t \in [0,1]$ be an interpolation parameter. Define the K-dimensional vector:

$$\underline{S}_{t,\mu} \equiv \sqrt{1-t/n} \sum_{i=1}^n X_{\mu i} \underline{W}_i^* + \Big(\int_0^t q(v) dv \Big)^{1/2} \underline{V}_\mu + \Big(\int_0^t (\rho - q(v)) dv \Big)^{1/2} \underline{U}_\mu^* \tag{17}$$

in which $q(v) \in \mathcal{S}_K^{++}(\rho)$ is a matrix valued interpolation path to be "adapted" later on. We will interpolate towards two auxiliary problems related to those discussed in sec. 3:

$$\left\{ \begin{array}{rcll} Y_{t,\mu} & \sim & P_{\rm out}(\,\cdot\,|\,\underline{S}_{t,\mu})\,, & 1 \le \mu \le m, \\ \underline{Y}'_{t,i} & = & \sqrt{t} R^{1/2}\, \underline{W}_i^* + \underline{Z}'_i\,, & 1 \le i \le n, \end{array} \right. \tag{18}$$

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. We recall that in our notation the $*$-variables have to be retrieved, while the other random variables are assumed to be known. 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 $\underline{U}_\mu^*$ (it thus depends on the full matrix $w = (w_{il})_{i=1,l=1}^{n,K}$). For $t \in [0,1]$ we now introduce the *interpolating posterior*:

$$P_t(w, u \mid Y, Y', X, V) = \frac{1}{\mathcal{Z}_n(t)} \prod_{i=1}^n P_0(\underline{w}_i) \prod_{\mu=1}^m P_{\rm out}(Y_{t,\mu} \mid \underline{s}_{t,\mu}) \prod_{i=1}^n e^{-\frac{1}{2}\|\underline{Y}'_{t,i} - \sqrt{t} R^{1/2} \underline{w}_i\|_2^2} \tag{19}$$

with $\mathcal{Z}_n(t)$ the normalization factor equal to the numerator integrated over all components of w and u. The average free entropy at time t is by definition

$$f_n(t) \equiv \frac{1}{n} \mathbb{E} \ln \mathcal{Z}_n(t). \tag{20}$$

One easily verifies that

$$\left\{ \begin{array}{rcl} f_n(0) & = & f_n - \frac{K}{2}\,, \\ f_n(1) & = & \psi_{P_0}(R) + \frac{m}{n} \Psi_{P_{\rm out}}(\int_0^1 q(t) dt; \rho) - \frac{1}{2} {\rm Tr}(R\rho) - \frac{K}{2}\,. \end{array} \right. \tag{21}$$

We will relate these two extreme values through the fundamental theorem of calculus

$$f_n(0) = f_n(1) - \int_0^1 \frac{df_n(t)}{dt}. \tag{22}$$

The next step is to compute the free entropy variation along the interpolation path (see appendices sec. A):

Proposition 5.1 (Free entropy variation). *Denote by $\langle - \rangle_{n,t}$ the (Gibbs) expectation w.r.t. the interpolating posterior (19). Set $u_y(x) \equiv -\ln P_{\rm out}(y|x)$. For all $t \in [0,1]$*

$$\frac{df_n(t)}{dt} = -\frac{1}{2} \mathbb{E} \Big\langle {\rm Tr} \Big[\Big(\frac{1}{n} \sum_{\mu=1}^m \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) \nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - R \Big) (Q - q(t)) \Big] \Big\rangle_{n,t} \tag{23}$$

$$+ \frac{1}{2} {\rm Tr}(R(q(t) - \rho)) + o_n(1)\,, \tag{24}$$

where ∇ is the K-dimensional gradient with respect to the argument of $u_{Y_{t,\mu}}(\cdot)$, and $o_n(1) \to 0$ in the $n, m \to \infty$ limit uniformly in $t \in [0,1]$, $Q_{ll'} \equiv \sum_{i=1}^n W_{il}^* w_{il'}/n$ is a $K \times K$ overlap matrix .

A crucial step of the adaptive interpolation method is to show that the overlap matrix entries concentrate. In order to do this we must introduce a "small" perturbation of the interpolating problem by *adding* to the system a small K-dimensional Gaussian "side channel"

$$\underline{\widehat{Y}}_i = \epsilon^{1/2}\underline{W}_i^* + \underline{\widehat{Z}}_i \tag{25}$$

with $\epsilon \in \mathcal{S}_K^{++}$, $\underline{\widehat{Z}}_i \sim \mathcal{N}(0, I_{K\times K})$. Note that $\epsilon^{1/2}$ is a matrix square root. With this extra channel the posterior (19) must be multiplied by $\prod_{i=1}^n \exp(-\|\underline{\widehat{Y}}_i - \epsilon^{1/2}\underline{w}_i\|_2^2/2)$. The corresponding average free entropy and Gibbs expectation are denoted $f_{n,\epsilon}$ and $\langle - \rangle_{n,t,\epsilon}$. An easy argument shows that

$$|f_{n,\epsilon}(t) - f_n(t)| \leq \|\epsilon\|_{\mathrm{F}} \frac{S^2 K}{2} \tag{26}$$

for all $t \in [0,1]$, where $S > 0$ such that the support of P_0 is included in the sphere of radius S and $\|M\|_{\mathrm{F}}^2$ denotes the Frobenius norm. This small perturbation forces the overlap to concentrate around its mean (see appendices sec. A for more details):

Proposition 5.2 (Overlap concentration). *There exists a sequence of matrices $\mathcal{S}_K^{++} \ni (\epsilon_n)_{n\geq 1} \to (0)$ (the all-zeros matrix) s.t.*

$$\lim_{n\to\infty} \int_0^1 dt\, \mathbb{E}\big\langle \|Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon_n}\|_{\mathrm{F}}^2 \big\rangle_{n,t,\epsilon_n} = 0\,. \tag{27}$$

Note that since $(\epsilon_n)_{n\geq 1}$ converges to (0), as claimed before $f_{n,\epsilon_n}(t)$ and $f_n(t)$ have the same limit (provided it exists). The adaptive choice of the interpolation path is based on the following:

Proposition 5.3 (Optimal interpolation path). *For all $R \in \mathcal{S}_K^+$ the matrix differential equation $q(t) = \mathbb{E}\,\langle Q\rangle_{n,t,\epsilon_n}$ admits a unique solution $q_n^{(R)}(t)$ in $\mathcal{S}_K^+(\rho)$ and the mapping $R \in \mathcal{S}_K^+ \mapsto \int_0^1 q_n^{(R)}(v)dv$ is continuous.*

Proof: To prove this proposition one first notes that $\mathbb{E}\,\langle Q\rangle_{n,t,\epsilon_n}$ is a matrix-valued function of $(t, \int_0^t q(v)dv) \in \mathbb{R}\times\mathcal{S}_K$. So we have to solve a first order differential equation in the *finite dimensional vector space* $\mathcal{S}_K$ of $K\times K$ matrices. It is then not difficult to verify that $\mathbb{E}\,\langle Q\rangle_{n,t,\epsilon_n}$ is a bounded $\mathcal{C}^1$ function of $(\int_0^t q(v)dv, R)$, and thus the proposition follows from a direct application of the parametric Cauchy-Lipschitz theorem. Since $\mathbb{E}\langle Q\rangle_{n,t,\epsilon_n}$ and $\rho - \mathbb{E}\langle Q\rangle_{n,t,\epsilon_n}$ are positive matrices (see appendices sec. A for the argument) we also have $q(t) \in \mathcal{S}_K^+(\rho)$ which ends the proof. □

Now define

$$f_{\mathrm{RS}}(q,R) \equiv \psi_{P_0}(R) + \alpha\Psi_{\mathrm{out}}(q;\rho) - \mathrm{Tr}(Rq)/2 \tag{28}$$

and call it the *replica symmetric* (RS) potential; this is nothing else than the function in the bracket appearing in the replica formula (12). Using the optimal interpolating function of Proposition 5.3 allows to relate this RS potential and free entropy f_n. Indeed by Cauchy-Schwarz the square of the r.h.s. of (23) is bounded by

$$\int_0^1 dt\, \mathbb{E}\Big\langle \Big\| \Big(\frac{1}{n}\sum_{\mu=1}^m \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal} - R\Big\|_{\mathrm{F}}^2\Big\rangle_{n,t,\epsilon_n} \times \int_0^1 dt\, \mathbb{E}\big\langle \|Q - q_n^{(R)}(t)\|_{\mathrm{F}}^2\big\rangle_{n,t,\epsilon_n}\,.$$

We claim that this upper bound equals $o_n(1)$. Indeed: (a) the first factor is bounded (independently of t) because we supposed that P_{out} is generated by (11) with assumptions (H1), (H2), (H3) (see appendices sec. A for a proof) and; (b) the second factor goes to 0 when $n, m \to \infty$ by an application of Proposition 5.2 and Proposition 5.3. Putting this result together with (21), (22) and Proposition 5.1 we arrive at:

Proposition 5.4 (Fundamental identity). *Let $(R_n)_{n\geq 1} \in (\mathcal{S}_K^+)^{\mathbb{N}}$ be a bounded sequence. For each $n \in \mathbb{N}$, let $q_n^{(R_n)}$ be the unique solution of the matrix differential equation $q(t) = \mathbb{E}\,\langle Q\rangle_{n,t,\epsilon_n}$. Then*

$$f_n = f_{\mathrm{RS}}\Big(\int_0^1 q_n^{(R_n)}(v)dv, R_n\Big) + o_n(1). \tag{29}$$

End of proof of Theorem 3.1: First, from the proposition 5.4 we trivially deduce the lower bound:

$$\liminf_{n\to\infty} f_n \geq \sup_{R\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q, R). \tag{30}$$

We now turn our attention to the upper bound. Let $P = 2\alpha\|\nabla\Psi_{P_{\mathrm{out}}}(\rho;\rho)\|_2 I$ where I is the $K \times K$ identity matrix. The mapping $R \in \mathcal{S}_K^+ \mapsto \int_0^1 q_n^{(R)}(v)dv$ is continuous, consequently the map $R \mapsto 2\alpha\nabla\Psi_{P_{\mathrm{out}}}(\int_0^1 q_n^{(R)}(t)dt;\rho)$ from $S_K^+(P) \to S_K^+(P)$ is also continuous ($\nabla\Psi_{P_{\mathrm{out}}}$ denotes the derivative of $\Psi_{P_{\mathrm{out}}}$ w.r.t. its first argument, and can be shown to be continuous and bounded). By Brouwer's fixed-point theorem (since $\mathcal{S}_K^+(P)$ is convex and compact), there exists a fixed point $R_n^* = 2\alpha\nabla\Psi_{P_{\mathrm{out}}}(\int_0^1 q_n^{(R_n^*)}(t)dt;\rho)$. Proposition 5.4 then implies

$$f_n = f_{\mathrm{RS}}\Big(\int_0^1 q_n^{(R_n^*)}(t)dt, R_n^*\Big) + o_n(1). \tag{31}$$

We now remark that

$$f_{\mathrm{RS}}\Big(\int_0^1 q_n^{(R_n^*)}(t)dt, R_n^*\Big) = \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q, R_n^*). \tag{32}$$

Indeed, the function $g_{R_n^*} : q \in \mathcal{S}_K^+(\rho) \mapsto f_{\mathrm{RS}}(q, R_n^*) \in \mathbb{R}$ can be shown to be convex (appendices sec. A) and its q-derivative is $\nabla g_{R_n^*}(q) = \alpha\nabla\Psi_{P_{\mathrm{out}}}(q) - R_n^*/2$. Since $\nabla g_{R_n^*}(\int_0^1 q_n^{(R_n^*)}(t)dt) = 0$ by definition of R_n^*, and $\mathcal{S}_K^+(\rho)$ is convex and compact, the minimum of $g_{R_n^*}$ is necessarily achieved at $\int_0^1 q_n^{(R_n^*)}(t)dt$. Therefore

$$f_n = \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q, R_n^*) + o_n(1) \leq \sup_{R\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q, R) + o_n(1) \tag{33}$$

and thus

$$\limsup_{n\to\infty} f_n \leq \sup_{R\in\mathcal{S}_K^+} \inf_{q\in\mathcal{S}_K^+(\rho)} f_{\mathrm{RS}}(q, R) \tag{34}$$

which concludes the proof when combined with the lower bound above. ■

6 Perspectives

In this paper we revisit a model of two-layers neural network known as the committee machine in the teacher-student scenario that allows for explicit evaluation of the optimal learning errors. While this model has been discussed in early statistical physics literature using the non-rigorous replica method, we show here how these statements can be put on a mathematically rigorous basis, building on recent progress in proving the replica formulas.

Another contribution is the design of an approximate message passing algorithm (see [39] for a python implementation on GitHub) that efficiently achieves the Bayes-optimal learning error in the limit of large dimensions for a range of parameters out of the so-called hard phase that is associated with a first order phase

transition appearing in the model.

Finally, 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. Similar large computational gaps have been previously identified in many problems and we believe that its identification in a multi-layer neural network model makes it a very interesting candidate for further mathematical studies of the energy landscape in deep learning [40, 12]. Note that 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 case for the present model.

In this paper we focused on a two-layers neural network, but we note that the analysis and algorithm can be readily extended to a multi-layer setting, see [22], as long as the total number of hidden neurons stays of order one while the dimension of the data and the number of samples both grow at the same rate.

There are many possible extensions of the present work, which we hope will motivate revisiting the statistical physics approach to learning neural networks. An important open case, for instance, 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 $\alpha = \Theta(1)$, $K/n \to 0$ as $n \to \infty$. 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. Additional funding is acknowledged by FK from "Chaire de recherche sur les modèles et sciences des données", Fondation CFM pour la Recherche-ENS;

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] H. 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. Phase transitions, optimal errors and optimality of message-passing in generalized linear models. *arXiv preprint arXiv:1708.03395*, 2017.

[12] M. Baity-Jesi, L. Sagun, M. Geiger, S. Spigler, G. B. 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. *CoRR*, abs/1705.02780, 2017.

[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] 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.

[30] 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.

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

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

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] L. Sagun, V. U. Guney, G. B. Arous, and Y. LeCun. Explorations on high dimensional landscapes. *arXiv preprint arXiv:1412.6615*, 2014.

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

[42] 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.

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

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

[45] 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.

[46] 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.

A Proof details for Theorem 3.1

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*

$$\mathbb{E}\langle g(Y, X^{(1)}, \dots, X^{(k)})\rangle = \mathbb{E}\langle g(Y, X^{(1)}, \dots, X^{(k-1)}, X)\rangle\,. \tag{35}$$

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 Proposition 5.3.

Lemma A.2. *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

$$\mathbb{E}\langle Q\rangle = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \underline{W}_i^* \underline{w}_i^\intercal\rangle = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\langle \underline{w}_i\rangle\langle \underline{w}_i^\intercal\rangle \tag{36}$$

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

$$\begin{aligned}\mathbb{E}(\rho - \langle Q\rangle) &= \frac{1}{n}\sum_{i=1}^n \Big(\mathbb{E}[\underline{W}_i^*(\underline{W}_i^*)^\intercal] - \mathbb{E}\langle \underline{w}_i\rangle\langle \underline{w}_i^\intercal\rangle\Big) \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\underline{W}_i^* - \langle \underline{w}_i\rangle)((\underline{W}_i^*)^\intercal - \langle \underline{w}_i^\intercal\rangle)]\end{aligned} \tag{37}$$

where the last equality is proved by an application of the Nishimori identity again. This last expression is obviously in $\mathcal{S}_K^+$. ■

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 $\nabla\nabla^\intercal u_y(x)$ 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

$$\nabla\nabla^\intercal u_{Y_\mu}(\underline{x}) + \nabla u_{Y_\mu}(\underline{x})\nabla^\intercal u_{Y_\mu}(\underline{x}) = \frac{\nabla\nabla^\intercal P_{\text{out}}(Y_\mu|\underline{x})}{P_{\text{out}}(Y_\mu|\underline{x})}\,. \tag{38}$$

Finaly we will use the matrices $w \in \mathbb{R}^{n\times K}$, $u \in \mathbb{R}^{m\times K}$, $Y \in \mathbb{R}^m$, $Y' \in \mathbb{R}^{n\times K}$, $X \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{m\times K}$,

$W^* \in \mathbb{R}^{n\times K}$ and $U^* \in \mathbb{R}^{m\times K}$. Like in sec. 5 we adopt the convention that all underlined vectors are K dimensional. For example $\underline{u}_\mu$, $\underline{U}_\mu$, $\underline{V}_\mu$, $\underline{Y}'_i$ are all K-dimensional.

It is convenient to reformulate the expression of the interpolating free entropy $f_n(t)$ in the Hamiltonian language. We introduce an interpolating Hamiltonian:

$$\mathcal{H}_t(w,u;Y_t,Y'_t,X,V) \equiv -\sum_{\mu=1}^{m} u_{Y_{t,\mu}}(\underline{s}_{t,\mu}) + \frac{1}{2}\sum_{i=1}^{n} \|\underline{Y}'_i - \sqrt{t}R^{1/2}\,\underline{w}_i\|_2^2\,. \tag{39}$$

The average free entropy at time t reads

$$f_n(t) \equiv \frac{1}{n}\mathbb{E}\ln \int_{\mathbb{R}^{n\times K}} dP_0(w) \int_{\mathbb{R}^{m\times K}} \mathcal{D}u\, e^{-\mathcal{H}_t(w,u;Y_t,Y'_t,X,V)} \tag{40}$$

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

$$\begin{aligned} f_n(t) = \frac{1}{n}\mathbb{E}_{X,V} \int dY_t dY'_t dP_0(W^*)\mathcal{D}U^* e^{-\mathcal{H}_t(W^*,U;Y_t,Y'_t,X,V)} \\ \times \ln \int dP_0(w)\mathcal{D}u\, e^{-\mathcal{H}_t(w,u;Y_t,Y'_t,X,V)}. \end{aligned} \tag{41}$$

We begin with the proof of Proposition 5.1 which we recall for the convenience of the reader.

Proposition A.3 (Free entropy variation)**.** *Denote by $\langle - \rangle_{n,t}$ the (Gibbs) expectation w.r.t. the interpolating posterior in sec. 5. Set $u_y(\underline{x}) \equiv -\ln P_{\rm out}(y|\underline{x})$. For all $t \in [0,1]$*

$$\begin{aligned} \frac{df_n(t)}{dt} = &-\frac{1}{2}\mathbb{E}\Big\langle \mathrm{Tr}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - R\Big)\big(Q - q(t)\big)\Big]\Big\rangle_{n,t} \\ &+ \frac{1}{2}\mathrm{Tr}(R(q(t)-\rho)) + o_n(1)\,, \end{aligned} \tag{42}$$

where ∇ is the K-dimensional gradient with respect to the argument of $u_{Y_{t,\mu}}(\cdot)$, and $o_n(1) \to 0$ in the $n,m \to \infty$ limit uniformly in $t \in [0,1]$, and the $K \times K$ overlap matrix $Q_{ll'} \equiv \frac{1}{n}\sum_{i=1}^{n} W^*_{il}w_{il'}$.

Proof: We drop the t index for the measurements Y_t, Y'_t as they are dummy variables. We will first prove that for all $t \in (0,1)$

$$\begin{aligned} \frac{df_n(t)}{dt} = &-\frac{1}{2}\mathbb{E}\Big\langle \mathrm{Tr}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^{m} \nabla u_{Y_\mu}(\underline{s}_{t,\mu})\nabla u_{Y_\mu}(\underline{S}_{t,\mu})^\intercal - R\Big)\Big(\frac{1}{n}\sum_{i=1}^{n}\underline{W}^*_i\underline{w}_i^\intercal - q(t)\Big)\Big\rangle_t\Big] \\ &+ \mathrm{Tr}\frac{R(q(t)-\rho)}{2} - \frac{A_n}{2}\,, \end{aligned} \tag{43}$$

where

$$A_n = \mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{\sqrt{n}}\sum_{\mu=1}^{m}\frac{\nabla\nabla^\intercal P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}{P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\underline{W}^*_i(\underline{W}^*_i)^\intercal - \rho)\Big)\Big]\frac{1}{n}\ln \mathcal{Z}_n(t)\Big]\,. \tag{44}$$

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 t-derivative is given by

$$\begin{aligned}\frac{d}{dt}\mathcal{H}_t(W^*,U^*;Y,Y',X,V) &= -\sum_{\mu=1}^{m}\nabla^\intercal u_{Y_\mu}(\underline{S}_{t,\mu})\frac{d\underline{S}_{t,\mu}}{dt}\\ &\quad -\frac{1}{2}\frac{1}{\sqrt{t}}\sum_{i=1}^{n}(R^{1/2}\underline{W}_i^*)^\intercal(\underline{Y}_i'-\sqrt{t}R^{1/2}\underline{W}_i^*)\\ &= -\sum_{\mu=1}^{m}\mathrm{Tr}\Big[\frac{d\underline{S}_{t,\mu}}{dt}\nabla^\intercal u_{Y_\mu}(\underline{S}_{t,\mu})\Big]\\ &\quad -\frac{1}{2}\frac{1}{\sqrt{t}}\sum_{i=1}^{n}\mathrm{Tr}\Big[R^{1/2}(\underline{Y}_i'-\sqrt{t}R^{1/2}\underline{W}_i^*)\underline{W}_i^{*T}\Big] \end{aligned} \tag{45}$$

(where we used that R is symmetric). The derivative of the interpolating free entropy thus reads, for $0<t<1$,

$$\frac{df_n(t)}{dt} = -\underbrace{\frac{1}{n}\mathbb{E}\big[\frac{d}{dt}\mathcal{H}_t'(W^*,U^*;Y,Y',X,V)\ln\mathcal{Z}_n(t)\big]}_{T_1} - \underbrace{\frac{1}{n}\mathbb{E}\big\langle\mathcal{H}_t'(w,u;Y,Y',X,V)\big\rangle_t}_{T_2}. \tag{46}$$

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

$$T_2 = \frac{1}{n}\mathbb{E}\big\langle\frac{d}{dt}\mathcal{H}_t(w,u;Y,Y',X,V)\big\rangle_t = \frac{1}{n}\mathbb{E}\,\frac{d}{dt}\mathcal{H}_t(W^*,U^*;Y,Y',X,V) = 0\,. \tag{47}$$

We now compute T_1. This involves matrix derivatives which have to be done carefully. We first note that the matrix $\int_0^1 q(s)ds \in \mathcal{S}_K^{++}$ and therefore $(\int_0^1 q(s)ds)^{1/2}$, $(\int_0^1 q(s)ds)^{-1/2}$ are well defined. Then,

$$\begin{aligned}\mathbb{E}\Big[\mathrm{Tr}\Big[\frac{d\underline{S}_{t,\mu}}{dt}\nabla^\intercal u_{Y_\mu}(\underline{S}_{t,\mu})\Big]\ln\mathcal{Z}_n(t)\Big] &= \frac{1}{2}\mathbb{E}\Big[\mathrm{Tr}\Big[\Big(-\frac{\sum_{i=1}^n X_{\mu i}\underline{W}_i^*}{\sqrt{n(1-t)}}\\ &+\frac{d}{dt}\Big(\int_0^t q(s)ds\Big)^{1/2}\underline{V}_\mu + \frac{d}{dt}\Big(\int_0^t(\rho-q(s))ds\Big)^{1/2}\underline{U}_\mu^*\Big)\nabla^\intercal u_{Y_\mu}(\underline{S}_{t,\mu})\Big]\ln\mathcal{Z}_n(t)\Big]. \end{aligned} \tag{48}$$

We then compute the first line of the right-hand side of (48). By Gaussian integration by parts w.r.t. $X_{\mu i}$ (recall hypothesis (H3)), and using the identity (38), we find after some algebra

$$\begin{aligned}&\frac{1}{\sqrt{n(1-t)}}\mathbb{E}\big[\mathrm{Tr}\Big[\sum_{i=1}^n X_{\mu i}\underline{W}_i^*\nabla^\intercal u_{Y_\mu}(\underline{S}_{t,\mu})\Big]\ln\mathcal{Z}_n(t)\big]\\ &\quad= \mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{n}\sum_{i=1}^n\underline{W}_i^*\underline{W}_i^\intercal\frac{\nabla\nabla^\intercal P_{\mathrm{out}}(Y_\mu|\underline{S}_{t,\mu})}{P_{\mathrm{out}}(Y_\mu|\underline{S}_{t,\mu})}\Big]\ln\mathcal{Z}_n(t)\Big]\\ &\qquad+\mathbb{E}\Big\langle\mathrm{Tr}\Big[\frac{1}{n}\sum_{i=1}^n\underline{W}_i^*\underline{w}_i^\intercal\nabla u_{Y_\mu}(\underline{S}_{t,\mu})\nabla^\intercal u_{Y_\mu}(\underline{s}_{t,\mu})\Big]\Big\rangle_t. \end{aligned} \tag{49}$$

Similarly for the second line of the right hand side of (48), 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

$$\frac{d}{dt}M(t) = (M(t))^{1/2}\frac{d(M(t))^{1/2}}{dt} + \frac{d(M(t))^{1/2}}{dt}(M(t))^{1/2} \tag{50}$$

for any $M(t)\in\mathcal{S}_K^+$, and the cyclicity and linearity of the trace. Applying (50) to $M(t)$ equal to $\int_0^t q(s)ds$ and

$\int_0^t(\rho - q(s))ds$, as well as the identity (38), we reach after some algebra

$$\mathbb{E}\Big[\mathrm{Tr}\Big[\Big(\frac{d}{dt}\Big(\int_0^t q(s)ds\Big)^{1/2}\underline{V}_\mu + \frac{d}{dt}\Big(\int_0^t(\rho-q(s))ds\Big)^{1/2}\underline{U}^*_\mu\Big)\nabla^\intercal u_{Y_\mu}(\underline{S}_{\mu,t})\Big]\ln\mathcal{Z}_n(t)\Big]$$
$$= \mathbb{E}\Big[\mathrm{Tr}\Big[\rho\frac{\nabla\nabla^\intercal P_{\rm out}(Y_\mu|\underline{S}_{\mu,t})}{P_{\rm out}(Y_\mu|\underline{S}_{\mu,t})}\Big]\ln\mathcal{Z}_n(t)\Big] + \mathbb{E}\Big\langle\mathrm{Tr}\Big[q(t)\nabla u_{Y_\mu}(\underline{S}_{\mu,t})\nabla^\intercal u_{Y_\mu}(\underline{s}_{\mu,t})\Big]\Big\rangle_t . \tag{51}$$

As seen from (45), (46) it remains to compute $\mathbb{E}[\mathrm{Tr}[R^{1/2}(\underline{Y}'_i - \sqrt{t}R^{1/2}\underline{W}^*_i)\underline{W}^{*T}_i]\ln\mathcal{Z}_n(t)]$. Recall that $\underline{Y}'_i - \sqrt{t}R^{1/2}\underline{W}^*_i = \underline{Z}'_i \sim \mathcal{N}(0,1)$. Using an integration by parts leads to

$$\mathbb{E}\Big[\mathrm{Tr}\Big[R^{1/2}(\underline{Y}'_i - \sqrt{t}R^{1/2}\underline{W}^{*T}_i)\Big]\ln\mathcal{Z}_n(t)\Big] = -\sqrt{t}\mathrm{Tr}\Big[R^{1/2}\big(\rho - \mathbb{E}\langle W^*_j w_j\rangle_t\big)\Big] . \tag{52}$$

Finaly the term T_1 is obtained by putting together (48), (49), (51) and (52).

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

$$\mathbb{E}\Big[\frac{\nabla\nabla^\intercal P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}{P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}\,\Big|\,W^*, \{\underline{S}_{t,\mu}\}_{\mu=1}^m\Big] = \int dY_\mu \nabla\nabla^\intercal P_{\rm out}(Y_\mu|\underline{S}_{t,\mu}) = 0 , \tag{53}$$

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

$$\mathbb{E}\Big[\mathrm{Tr}\Big[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{\nabla\nabla^\intercal P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}{P_{\rm out}(Y_\mu|\underline{S}_{t,\mu})}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n(\underline{W}^*_i(\underline{W}^*_i)^\intercal - \rho)\Big)\Big]\Big] = 0 \tag{54}$$

Next, one shows by standard second moment methods that $\mathbb{E}[(n^{-1}\ln\mathcal{Z}_n(t) - f_n(t))^2] \to 0$ as $n \to +\infty$ uniformly in $t \in [0,1]$. Then, using this last fact together with (54), under hypothesis $(H1), (H2), (H3)$ 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 of formula (42) for the free entropy variation. ■

We now turn to the proof of Proposition 5.2 which we restate here:

Proposition A.4 (Overlap concentration). *There exists a sequence of $K \times K$ matrices $(\epsilon_n)_{n\geq 1} \in \mathcal{S}_K^{++}$ that converges to the all-zeros matrix (0) such that*

$$\lim_{n\to\infty}\int_0^1 dt\,\mathbb{E}\big\langle\|Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon_n}\|_{\rm F}^2\big\rangle_{n,t,\epsilon_n} = 0 . \tag{55}$$

Proof: Recall the perturbation (or side information channel) added to the posterior. We take a $K \times K$ matrix $\epsilon \in \mathcal{S}_K^{++}$ and denote the matrix elements by $\epsilon_{ll'}$. Recall that we add a K-dimensional Gaussian side channel $\widehat{\underline{Y}}_i = \epsilon^{1/2}\underline{W}^*_i + \widehat{\underline{Z}}_i$ with i.i.d. $\widehat{Z}_{il} \sim \mathcal{N}(0,1)$ (here $i = 1,\dots,n$ and $l = 1,\dots,K$). Note that here $\epsilon^{1/2}$ is the matrix square root of ϵ. This multiplies the posterior by a term

$$\prod_{i=1}^n e^{-\frac{1}{2}\|\widehat{\underline{Y}}_i - \epsilon^{1/2}\underline{w}_i\|_2^2} \tag{56}$$

or equivalently adds to the Hamiltonian (39) a term (we remark that since ϵ and $\epsilon^{1/2}$ are symmetric we have $\underline{w}_i^\intercal\epsilon\underline{W}^*_i = (\underline{W}^*_i)^\intercal\epsilon\underline{w}_i$ and $\underline{w}_i^\intercal\epsilon^{1/2}\,\underline{\widehat{Z}}_i = \underline{\widehat{Z}}_i^\intercal\epsilon^{1/2}\,\underline{w}_i$)

$$\mathcal{H}_{\rm pert} \equiv \sum_{i=1}^n\Big(\frac{1}{2}\underline{w}_i^\intercal\epsilon\underline{w}_i - \underline{w}_i^\intercal\epsilon\,\underline{W}^*_i - \underline{w}_i^\intercal\epsilon^{1/2}\,\underline{\widehat{Z}}_i\Big). \tag{57}$$

In [24, 11] we show how to prove concentration for the case of a scalar side channel and the proof is generic as long as the side channel is added to a generic Hamiltonian (here (39)) which comes from a Bayes-optimal inference problem and thus satisfies the Nishimori identities in Proposition A.1. The proof here is conceptually similar, except that one has to look at the effect of the perturbation in all "directions", i.e., with respect to each separate variation of the matrix elements $\epsilon_{ll'} = \epsilon_{l'l}$. In particular the derivative of the added Hamiltonian with respect to one matrix element, that must remain symmetric, then reads

$$\frac{d}{d\epsilon_{ll'}}\mathcal{H}_{\text{pert}} \equiv n\mathcal{L}_{ll'} = \frac{1}{2}\sum_{i=1}^{n}\Big(w_{il}w_{il'} - w_{il}W^*_{il'} - w_{il'}W^*_{il} - 2\underline{w}_i^\intercal \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big). \tag{58}$$

After some lenghty algebra (similar to [24], see sec. B of the appendices for the details) using Gaussian integration by parts and the Nishimori identity one obtains the following fluctuation identity:

$$\begin{aligned}\mathbb{E}\big[\big\langle(\mathcal{L}_{ll'} - \mathbb{E}\langle\mathcal{L}_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n}\big] + C\Big\{\mathbb{E}\big\langle(\mathcal{L}_{l'l'} - \langle\mathcal{L}_{l'l'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n}\mathbb{E}\big\langle(\mathcal{L}_{ll} - \langle\mathcal{L}_{ll}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n}\Big\}^{1/4}\\ \geq \frac{1}{4}\mathbb{E}\big\langle(Q_{ll'} - \mathbb{E}\langle Q_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n} + \mathcal{O}(K^3/n)\end{aligned} \tag{59}$$

Therefore, in order to control the overlap fluctuations one needs to control those of $\mathcal{L}_{ll'}$. Fortunately this can be done. The proof found in [24] of the following lemma applies verbatim (to all elements $\mathcal{L}_{ll'}$ independently):

Lemma A.5 (Concentration of $\mathcal{L}_{ll'}$)**.** *There exists a sequence of $K \times K$ matrices $(\epsilon_n)_{n\geq 1} \in \mathcal{S}_K^{++}$ that converges to the all-zeros matrix (0) such that for all $l, l' \in \{1, \dots, K\}$*

$$\lim_{n\to\infty}\int_0^1 dt\,\mathbb{E}\big\langle(\mathcal{L}_{ll'} - \mathbb{E}\langle\mathcal{L}_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n} = 0. \tag{60}$$

As a consequence the following statement is also true under the same conditions:

$$\lim_{n\to\infty}\int_0^1 dt\,\mathbb{E}\big\langle(\mathcal{L}_{ll'} - \langle\mathcal{L}_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n} = 0. \tag{61}$$

The concentration of $Q_{ll'}$ then follows from the one of $\mathcal{L}_{ll'}$ as we explain now. From (59) combined with Cauchy-Schwarz we have for some constant $C' > 0$

$$\begin{aligned}\int_0^1 dt\,\mathbb{E}\big\langle(Q_{ll'} - \mathbb{E}\langle Q_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n} \leq 4\int_0^1 dt\,\mathbb{E}\big\langle(\mathcal{L}_{ll'} - \mathbb{E}\langle\mathcal{L}_{ll'}\rangle_{n,t,\epsilon_n})^2\big\rangle_{n,t,\epsilon_n}\\ + C'\Big\{\int_0^1 dt\,\mathbb{E}\big\langle(\mathcal{L}_{l'l'} - \langle\mathcal{L}_{l'l'}\rangle)^2\big\rangle\int_0^1 dt\,\mathbb{E}\big\langle(\mathcal{L}_{ll} - \langle\mathcal{L}_{ll}\rangle)^2\big\rangle\Big\}^{1/4} + \mathcal{O}(K^3/n).\end{aligned} \tag{62}$$

Taking the limit $n \to +\infty$ of this inequality, applying Lemma A.5 and then summing the resulting $K^2 = \mathcal{O}(1)$ fluctuations, we obtain the claimed result (55). ∎

Lemma A.6 (Boundedness of an overlap fluctuation)**.** *Under hypothesis (H2) one can find a constant $C(\varphi, K, \Delta) < +\infty$ (independent of n, t, ϵ_n) such that for any $R_n \in \mathcal{S}_K^+$ we have*

$$\mathbb{E}\Big\langle\Big\|\frac{1}{n}\sum_{\mu=1}^{m}\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^\intercal - R_n\Big\|_{\text{F}}^2\Big\rangle_{n,t,\epsilon_n} \leq 2\text{Tr}(R_n^2) + \alpha^2 C(\varphi, K, \Delta). \tag{63}$$

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 $\mathrm{Tr}(A-B)^2 \le 2(\mathrm{Tr}A^2+\mathrm{Tr}B^2)$. Therefore

$$\mathbb{E}\Big\langle \|\frac{1}{n}\sum_{\mu=1}^{m}\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal} - R_n\|_{\mathrm{F}}^2\Big\rangle_{n,t,\epsilon_n}$$
$$\le 2\mathrm{Tr}(R_n^2) + 2\mathbb{E}\Big\langle \mathrm{Tr}\Big(\frac{1}{n}\sum_{\mu=1}^{m}\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal}\Big)^2\Big\rangle_{n,t,\epsilon_n}. \tag{64}$$

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

$$\mathbb{E}\Big\langle \|\frac{1}{n}\sum_{\mu=1}^{m}\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal}\|_{\mathrm{F}}^2\Big\rangle_{n,t,\epsilon_n} \le \mathbb{E}\Big\langle \frac{1}{n^2}\Big(\sum_{\mu=1}^{m}\|\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal}\|_{\mathrm{F}}\Big)^2\Big\rangle_{n,t,\epsilon_n}$$
$$\le \mathbb{E}\Big\langle \frac{1}{n^2}\Big(\sum_{\mu=1}^{m}\|\nabla u_{Y_{t,\mu}}(\underline{s}_{t,\mu})\|_2\|\nabla u_{Y_{t,\mu}}(\underline{S}_{t,\mu})^{\intercal}\|_2\Big)^2\Big\rangle_{n,t,\epsilon_n}. \tag{65}$$

From the random representation of the transition kernel,

$$u_{Y_{t,\mu}}(\underline{s}) = \ln P_{\mathrm{out}}(Y_{t,\mu}|\underline{x}) = \ln\int dP_A(a_\mu)\frac{1}{\sqrt{2\pi\Delta}}e^{-\frac{1}{2\Delta}(Y_{t,\mu}-\varphi(\underline{x},a_\mu))^2} \tag{66}$$

and thus

$$\nabla u_{Y_{t,\mu}}(\underline{x}) = \frac{\int dP_A(a_\mu)(Y_{t,\mu}-\varphi(\underline{x},a_\mu))\nabla\varphi(\underline{x},a_\mu)e^{-\frac{1}{2\Delta}(Y_{t,\mu}-\varphi(\underline{x},a_\mu))^2}}{\int dP_A(a_\mu)e^{-\frac{1}{2\Delta}(Y_{t,\mu}-\varphi(\underline{x},a_\mu))^2}} \tag{67}$$

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 $|Y_{t,\mu}| \le \sup|\varphi| + \sqrt{\Delta}|Z_\mu|$, where the supremum is taken over both arguments of φ, and thus we immediately obtain for all $\underline{s}\in\mathbb{R}^K$

$$\|\nabla u_{Y_{t,\mu}}(\underline{x})\| \le (2\sup|\varphi| + \sqrt{\Delta}|Z_\mu|)\sup\|\nabla\varphi\|. \tag{68}$$

From (68) and (65) we see that it suffices to check that

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

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.6. ■

Lemma A.7 (Convexity of $\Psi_{P_{\mathrm{out}}}$). *Recall that $\Psi_{P_{\mathrm{out}}}$ is defined as the free entropy of the second auxiliary channel (7). More precisely, for $q\in\mathcal{S}_K^+(\rho)$, we have:*

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

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

Proof: The continuity and differentiability of $\Psi_{P_{\mathrm{out}}}$ is easy, and exactly similar to the first part of the proof of Proposition 11 of [11]. 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 $u_Y(x) \equiv \ln P_{\rm out}(y|x)$. We define the average $\langle - \rangle_{\rm sc}$ as

$$\langle g(w) \rangle_{\rm sc} \equiv \frac{\int_{\mathbb{R}^K} \mathcal{D}w P_{\rm out}\left((\rho-q)^{1/2}w + q^{1/2}V\right) g(w)}{\int_{\mathbb{R}^K} \mathcal{D}w P_{\rm out}\left((\rho-q)^{1/2}w + q^{1/2}V\right)}, \tag{69}$$

for any continuous bounded function g. One arrives at:

$$\nabla \Psi_{P_{\rm out}}(q) = \frac{1}{2}\mathbb{E}\Big\langle \nabla u_Y\left((\rho-q)^{1/2}W^* + q^{1/2}V\right) \nabla u_Y\left((\rho-q)^{1/2}w + q^{1/2}V\right)^\intercal \Big\rangle_{\rm sc}. \tag{70}$$

Note that this gradient is actually a matrix of size $K \times K$, as it is a gradient w.r.t. $q \in \mathbb{R}^{K\times K}$. The Hessian of $\Psi_{P_{\rm out}}$ w.r.t. q is thus a 4-tensor. One can compute in the same way:

$$\begin{aligned}\nabla^2 \Psi_{P_{\rm out}}(q) = \frac{1}{2}\mathbb{E}\Big[\Big(\Big\langle \frac{\nabla^2 P_{\rm out}\left((\rho-q)^{1/2}w + q^{1/2}V\right)}{P_{\rm out}\left((\rho-q)^{1/2}w + q^{1/2}V\right)}\Big\rangle_{\rm sc} \\ - \Big\langle \nabla u_Y\left((\rho-q)^{1/2}W^* + q^{1/2}V\right) \nabla u_Y\left((\rho-q)^{1/2}w + q^{1/2}V\right)^\intercal \Big\rangle_{\rm sc}\Big)^{\otimes 2}\Big].\end{aligned} \tag{71}$$

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 $\Psi_{P_{\rm out}}$ is always positive, when seen as a matrix with rows and columns in $\mathbb{R}^{K\times K}$, and thus $\Psi_{P_{\rm out}}$ is convex, which ends the proof of Lemma A.7. ■

B A fluctuation identity

In this section we drop the indices in the Gibbs bracket that will simply be written as $\langle - \rangle$ as these do not play any role in the following analysis. We will relate the fluctuations of the object (58) that appears naturally in the problem and for which we can control its fluctuation that we recall here,

$$\mathcal{L}_{ll'} = \frac{1}{2n}\sum_{i=1}^n \left(w_{il}w_{il'} - w_{il}W^*_{il'} - w_{il'}W^*_{il} - 2\underline{w}_i^\intercal \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_i\right) \tag{72}$$

to those of each matrix element of the overlap matrix $Q_{ll'} = \frac{1}{n}\sum_{i=1}^n W^*_{il}w_{il'}$, namely we will prove the general fluctuation identity:

$$\begin{aligned}\mathbb{E}\big[\big\langle (\mathcal{L}_{ll'} - \mathbb{E}\langle \mathcal{L}_{ll'}\rangle)^2\big\rangle\big] &= \mathbb{E}\big[\big\langle (\mathcal{L}_{l'l} - \mathbb{E}\langle \mathcal{L}_{l'l}\rangle)^2\big\rangle\big] \\ &= \frac{1}{2}\Big\{\mathbb{E}\langle Q_{ll'}Q_{l'l}\rangle - \mathbb{E}[\langle Q_{ll'}\rangle\langle Q_{l'l}\rangle]\Big\} + \frac{1}{4}\Big\{\mathbb{E}\langle Q^2_{ll'}\rangle - \mathbb{E}[\langle Q_{ll'}\rangle]^2\Big\} + \mathcal{O}(K^3/n) \\ &= \frac{1}{2}\Big\{\mathbb{E}\langle Q_{ll'}Q_{l'l}\rangle - \mathbb{E}[\langle Q_{ll'}\rangle\langle Q_{l'l}\rangle]\Big\} + \frac{1}{4}\Big\{\mathbb{E}\langle Q^2_{l'l}\rangle - \mathbb{E}[\langle Q_{l'l}\rangle]^2\Big\} + \mathcal{O}(K^3/n).\end{aligned} \tag{73}$$

Identity (73) follows from summing the two following identities that we prove next:

$$\mathbb{E}[\langle\mathcal{L}_{ll'}^2\rangle] - \mathbb{E}[\langle\mathcal{L}_{ll'}\rangle^2] = \frac{1}{4}\mathbb{E}\Big[\langle Q_{ll'}^2\rangle - \big(\langle Q_{ll'}\rangle + \langle Q_{l'l}\rangle\big)^2 + \langle Q_{ll'}Q_{l'l}\rangle + \frac{2}{n^2}\sum_{i,j=1}^n \langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle\Big] + \mathcal{O}(K^2/n), \tag{74}$$

$$\mathbb{E}[\langle\mathcal{L}_{ll'}\rangle^2] - \mathbb{E}[\langle\mathcal{L}_{ll'}\rangle]^2 = \frac{1}{4}\mathbb{E}\Big[\langle Q_{l'l}\rangle^2 + \langle Q_{ll'}\rangle^2 + \langle Q_{ll'}Q_{l'l}\rangle - \frac{2}{n^2}\sum_{i,j=1}^n \langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle\Big] - \frac{1}{4}\mathbb{E}[\langle Q_{l'l}\rangle]^2 + \mathcal{O}(K^3/n), \tag{75}$$

From (73) we finally derive (59) that we recall here:

$$\mathbb{E}\big[\big\langle(\mathcal{L}_{ll'} - \mathbb{E}\langle\mathcal{L}_{ll'}\rangle)^2\big\rangle\big] + C\big\{\mathbb{E}\big\langle(\mathcal{L}_{l'l'} - \langle\mathcal{L}_{l'l'}\rangle)^2\big\rangle\mathbb{E}\big\langle(\mathcal{L}_{ll} - \langle\mathcal{L}_{ll}\rangle)^2\big\rangle\big\}^{1/4} \geq \frac{1}{4}\mathbb{E}\big\langle(Q_{ll'} - \mathbb{E}\langle Q_{ll'}\rangle)^2\big\rangle + \mathcal{O}(K^3/n) = \frac{1}{4}\mathbb{E}\big\langle(Q_{l'l} - \mathbb{E}\langle Q_{l'l}\rangle)^2\big\rangle + \mathcal{O}(K^3/n) \tag{76}$$

by showing in the last section of this appendix that for some constant $C > 0$,

$$\mathbb{E}\big\langle(Q_{ll'} - \langle Q_{ll'}\rangle)^2\big\rangle \leq C\,\mathbb{E}\big[\big\langle(\mathcal{L}_{l'l'} - \langle\mathcal{L}_{l'l'}\rangle)^2\big\rangle\big]^{1/2} \tag{77}$$

and similarly for $Q_{l'l}$, which direclty implies by Cauchy-Schwarz that

$$\big|\mathbb{E}\langle Q_{ll'}Q_{l'l}\rangle - \mathbb{E}[\langle Q_{ll'}\rangle\langle Q_{l'l}\rangle]\big| = \big|\mathbb{E}\big\langle(Q_{ll'} - \langle Q_{ll'}\rangle)(Q_{l'l} - \langle Q_{l'l}\rangle)\big\rangle\big| \leq \big\{\mathbb{E}\big\langle(Q_{ll'} - \langle Q_{ll'}\rangle)^2\big\rangle\mathbb{E}\big\langle(Q_{l'l} - \langle Q_{l'l}\rangle)^2\big\rangle\big\}^{1/2} \leq C\big\{\mathbb{E}\big\langle(\mathcal{L}_{l'l'} - \langle\mathcal{L}_{l'l'}\rangle)^2\big\rangle\mathbb{E}\big\langle(\mathcal{L}_{ll} - \langle\mathcal{L}_{ll}\rangle)^2\big\rangle\big\}^{1/4}. \tag{78}$$

The derivation of all these identities is lenghty but do not present any technical difficulty, and is a generalization of the proof of the fluctuation identity in [24] which is for the scalar case $K = 1$.

B.1 Preliminaries

We start with some preliminary computations that will be useful in the derivation of the two above identities. First we compute

$$\mathbb{E}[\langle\mathcal{L}_{ll'}\rangle] = \mathbb{E}\Big\langle\frac{1}{2n}\sum_{i=1}^n\Big(w_{il}w_{il'} - w_{il}W_{il'}^* - w_{il'}W_{il}^* - 2\underline{w}_i^\intercal\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_i\Big)\Big\rangle \overset{\text{N}}{=} \mathbb{E}\frac{1}{n}\sum_{i=1}^n\Big(\frac{1}{2}\langle w_{il}w_{il'}\rangle - \langle w_{il}\rangle\langle w_{il'}\rangle - \langle\underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_i\Big), \tag{79}$$

where here we used the Nishimori Proposition A.1 which in this case reads $\mathbb{E}[\langle w_{il}\rangle W_{il'}^*] = \mathbb{E}[\langle w_{il}\rangle\langle w_{il'}\rangle]$. Each time we use an identity that is a consequence of Proposition A.1 we write a N on top of the equality (that stands for "Nishimori"). In order to simplify this expression, we show a Gaussian integration by part mechanism that we will use repedately. We want to integrate by part the Gaussian noise in a term of the form (introducing a multiplicative $A = A(\underline{w}_i)$ term, that does not depend explicitely on the noise, in order to obtain

a more general identity that will be useful later on)

$$\begin{aligned}\mathbb{E}\Big[\langle A\underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_i\Big] &= \sum_{k,k'=1}^{K}\mathbb{E}\Big[\langle Aw_{ik}\rangle\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\widehat{Z}_{ik'}\Big] = \sum_{k,k'=1}^{K}\mathbb{E}\Big[\frac{d\langle Aw_{ik}\rangle}{d\widehat{Z}_{ik'}}\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\Big]\\ &= \sum_{k,k'=1}^{K}\mathbb{E}\Big[\big(\langle Aw_{ik}(\epsilon^{1/2}\underline{w}_i)_{k'}\rangle - \langle Aw_{ik}\rangle(\epsilon^{1/2}\langle\underline{w}_i\rangle)_{k'}\big)\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\Big]\\ &= \mathbb{E}\Big[\Big\langle A\underline{w}_i^\intercal\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\epsilon^{1/2}\underline{w}_i\Big\rangle - \langle A\underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\epsilon^{1/2}\langle\underline{w}_i\rangle\Big],\end{aligned} \tag{80}$$

where we used the Gaussian integration by part formula (or Stein lemma) $\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)]$ for $Z\sim\mathcal{N}(0,1)$, together with the fact that the derivative of the perturbing Hamiltonian is

$$\frac{d\mathcal{H}_{\text{pert}}}{d\widehat{Z}_{ik}} = -(\epsilon^{1/2}\underline{w}_i)_k, \quad \text{and thus} \quad \frac{d\langle\cdot\rangle}{d\widehat{Z}_{ik}} = \langle\cdot\,(\epsilon^{1/2}\underline{w}_i)_k\rangle - \langle\cdot\rangle(\epsilon^{1/2}\langle\underline{w}_i\rangle)_k\,. \tag{81}$$

Now we can write in general (we use $\underline{w}_i^{(b)}$ and $\underline{w}_i^{(c)}$ with $b,c\in\{1,2\}$ to distinguish i.i.d. replicas with product measure $\langle-\rangle^{\otimes 2}$ if $b\neq c$ and common one else, and denote $A^{(a)}$, $a\in\{1,2,3,\dots\}$, to emphasize that A depends on replica $\underline{w}_i^{(a)}$)

$$\begin{aligned}A^{(a)}(\underline{w}_i^{(b)})^\intercal\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\epsilon^{1/2}\underline{w}_i^{(c)} &= \frac{1}{2}A^{(a)}(\underline{w}_i^{(b)})^\intercal\Big[\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\epsilon^{1/2}+\epsilon^{1/2}\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big]\underline{w}_i^{(c)}\\ &= \frac{1}{2}A^{(a)}(\underline{w}_i^{(b)})^\intercal\frac{d\epsilon}{d\epsilon_{ll'}}\underline{w}_i^{(c)} = \frac{1}{2}A^{(a)}w_{il}^{(b)}w_{il'}^{(c)} = \frac{1}{2}A^{(a)}w_{il'}^{(b)}w_{il}^{(c)}\end{aligned} \tag{82}$$

where we used the following formula

$$\epsilon^{1/2}\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}} + \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\epsilon^{1/2} = \frac{d\epsilon}{d\epsilon_{ll'}} \tag{83}$$

and that the matrices are symmetric and thus $\frac{d\epsilon}{d\epsilon_{ll'}} = \frac{d\epsilon}{d\epsilon_{l'l}}$. Thus (80) becomes

$$\mathbb{E}\Big[\langle A\underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_i\Big] = \frac{1}{2}\big(\langle Aw_{il}w_{il'}\rangle - \langle Aw_{il}\rangle\langle w_{il'}\rangle\big) = \frac{1}{2}\big(\langle Aw_{il}w_{il'}\rangle - \langle Aw_{il'}\rangle\langle w_{il}\rangle\big). \tag{84}$$

Using this (79) becomes

$$\mathbb{E}[\langle\mathcal{L}_{ll'}\rangle] = -\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\big[\langle w_{il}\rangle\langle w_{il'}\rangle\big] \overset{\text{N}}{=} -\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\big[W_{il}^*\langle w_{il'}\rangle\big] = -\frac{1}{2}\mathbb{E}\langle Q_{ll'}\rangle = -\frac{1}{2}\mathbb{E}\langle Q_{l'l}\rangle. \tag{85}$$

We will need a further generalization of (84), where now the integration by part is done over three distinct terms (here A, B do not explicitely depend on the noise, only the Gibbs brackets $\langle A\rangle$, $\langle B\rangle$ do):

$$\begin{aligned}\mathbb{E}\Big[\langle A\rangle\langle B\rangle\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\,\widehat{\underline{Z}}_i\Big] &= \sum_{k,k'=1}^{K}\mathbb{E}\Big[\langle A\rangle\langle B\rangle\frac{d\langle w_{ik}\rangle}{d\widehat{Z}_{ik'}}\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'} + \langle w_{ik}\rangle\frac{d\langle A\rangle}{d\widehat{Z}_{ik'}}\langle B\rangle\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\\ &\qquad + \langle w_{ik}\rangle\langle A\rangle\frac{d\langle B\rangle}{d\widehat{Z}_{ik'}}\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\Big]\\ &\overset{(81)}{=} \sum_{k,k'=1}^{K}\mathbb{E}\Big[\langle A\rangle\langle B\rangle\big(\langle w_{ik}(\epsilon^{1/2}\underline{w}_i)_{k'}\rangle - \langle w_{ik}\rangle(\epsilon^{1/2}\langle\underline{w}_i\rangle)_{k'}\big)\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\\ &\qquad + \langle w_{ik}\rangle\langle B\rangle\big(\langle A(\epsilon^{1/2}\underline{w}_i)_{k'}\rangle - \langle A\rangle(\epsilon^{1/2}\langle\underline{w}_i\rangle)_{k'}\big)\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\\ &\qquad + \langle w_{ik}\rangle\langle A\rangle\big(\langle B(\epsilon^{1/2}\underline{w}_i)_{k'}\rangle - \langle B\rangle(\epsilon^{1/2}\langle\underline{w}_i\rangle)_{k'}\big)\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\Big].\end{aligned} \tag{86}$$

Thus, using (82) with $A^{(a)}\to A^{(a)}B^{(a')}$ we obtain after simplification

$$\begin{aligned}\mathbb{E}\Big[\langle A\rangle\langle B\rangle\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\,\widehat{\underline{Z}}_i\Big] &= \frac{1}{2}\mathbb{E}\Big[\langle A\rangle\langle B\rangle\langle w_{il}w_{il'}\rangle - 3\langle A\rangle\langle B\rangle\langle w_{il}\rangle\langle w_{il'}\rangle\\ &\qquad + \langle w_{il}\rangle\langle B\rangle\langle Aw_{il'}\rangle + \langle w_{il}\rangle\langle A\rangle\langle Bw_{il'}\rangle\Big]\end{aligned} \tag{87}$$

and in particular

$$\mathbb{E}\Big[\langle A\rangle\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\,\widehat{\underline{Z}}_i\Big] = \frac{1}{2}\mathbb{E}\Big[\langle A\rangle\langle w_{il}w_{il'}\rangle - 2\langle A\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + \langle w_{il}\rangle\langle Aw_{il'}\rangle\Big]. \tag{88}$$

A last required formula of the same type, derived similarly, is

$$\mathbb{E}\Big[\langle A\rangle\langle B\underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\,\widehat{\underline{Z}}_i\Big] = \frac{1}{2}\mathbb{E}\Big[\langle A\rangle\langle Bw_{il}w_{il'}\rangle - 2\langle A\rangle\langle Bw_{il}\rangle\langle w_{il'}\rangle + \langle Bw_{il}\rangle\langle Aw_{il'}\rangle\Big]. \tag{89}$$

Finally we will need the following overlap identities:

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}\big[\langle w_{il}w_{jl'}\rangle\langle w_{ik}w_{jk'}\rangle\big] \overset{\text{N}}{=} \mathbb{E}\Big\langle\frac{1}{n}\sum_{i=1}^{n}W_{il}^*w_{ik}\frac{1}{n}\sum_{j=1}^{n}W_{jl'}^*w_{jk'}\Big\rangle = \mathbb{E}[\langle Q_{lk}Q_{l'k'}\rangle] = \mathbb{E}[\langle Q_{kl}Q_{k'l'}\rangle], \tag{90}$$

$$\frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}\big[\langle w_{il}\rangle\langle w_{jl'}\rangle\langle w_{ik}w_{jk'}\rangle\big] \overset{\text{N}}{=} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}W_{ik}^*\langle w_{il}\rangle\frac{1}{n}\sum_{j=1}^{n}W_{jk'}^*\langle w_{jl'}\rangle\Big] = \mathbb{E}[\langle Q_{kl}\rangle\langle Q_{k'l'}\rangle]. \tag{91}$$

B.2 Derivation of (74)

We start with the first identity, namely the "thermal" fluctuations. Recall (58). Acting with $n^{-1}d/d\epsilon_{ll'}$ on both sides of (85) we thus obtain

$$-\mathbb{E}[\langle\mathcal{L}_{ll'}^2\rangle - \langle\mathcal{L}_{ll'}\rangle^2] + \frac{1}{n}\mathbb{E}\Big[\Big\langle\frac{d\mathcal{L}_{ll'}}{d\epsilon_{ll'}}\Big\rangle\Big] = \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\big[W_{il}^*\big(\langle w_{il'}\mathcal{L}_{ll'}\rangle - \langle w_{il'}\rangle\langle\mathcal{L}_{ll'}\rangle\big)\big]. \tag{92}$$

Computing the derivative of $\mathcal{L}_{ll'}$ and using $-2(d\epsilon^{1/2}/d\epsilon_{ll'})^2 = \epsilon^{1/2}(d^2\epsilon^{1/2}/d\epsilon_{ll'}^2) + (d^2\epsilon^{1/2}/d\epsilon_{ll'}^2)\epsilon^{1/2}$ which follows from (83) we find

$$\frac{1}{n}\mathbb{E}\Big[\Big\langle\frac{d\mathcal{L}_{ll'}}{d\epsilon_{ll'}}\Big\rangle\Big] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\Big[\Big\langle\underline{w}_i^\intercal\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)^2\underline{w}_i\Big\rangle - \langle\underline{w}_i^\intercal\rangle\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)^2\langle\underline{w}_i\rangle\Big] = \mathcal{O}(K^2/n). \tag{93}$$

Let us compute the term in (92):

$$\begin{aligned}
&\mathbb{E}\big[W_{il}^*\big(\langle w_{il'}\mathcal{L}_{ll'}\rangle - \langle w_{il'}\rangle\langle\mathcal{L}_{ll'}\rangle\big)\big] \\
&\quad= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Big[\frac{1}{2}W_{il}^*\langle w_{il'}w_{jl}w_{jl'}\rangle - \frac{1}{2}W_{il}^*W_{jl'}^*\langle w_{il'}w_{jl}\rangle - \frac{1}{2}W_{il}^*W_{jl}^*\langle w_{il'}w_{jl'}\rangle - W_{il}^*\langle w_{il'}\underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j \\
&\quad- \frac{1}{2}W_{il}^*\langle w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle + \frac{1}{2}W_{il}^*W_{jl'}^*\langle w_{il'}\rangle\langle w_{jl}\rangle + \frac{1}{2}W_{il}^*W_{jl}^*\langle w_{il'}\rangle\langle w_{jl'}\rangle + W_{il}^*\langle w_{il'}\rangle\langle\underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big].
\end{aligned} \tag{94}$$

Let us integrate by parts the two terms involving the explicit noise dependence (it is important that this is done *before* employing the Nishimori identity):

$$-\mathbb{E}\Big[W_{il}^*\langle w_{il'}\underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big] \overset{(84)}{=} -\frac{1}{2}\mathbb{E}\Big[W_{il}^*\langle w_{il'}w_{jl'}w_{jl}\rangle - W_{il}^*\langle w_{il'}w_{jl}\rangle\langle w_{jl'}\rangle\Big], \tag{95}$$

$$\begin{aligned}
\mathbb{E}\Big[W_{il}^*\langle w_{il'}\rangle\langle\underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big] &\overset{(88)}{=} \frac{1}{2}\mathbb{E}\Big[W_{il}^*\langle w_{il'}\rangle\langle w_{jl'}w_{jl}\rangle - 2W_{il}^*\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle \\
&\qquad + W_{il}^*\langle w_{il'}w_{jl'}\rangle\langle w_{jl}\rangle\Big].
\end{aligned} \tag{96}$$

Plugging these results in (94), (92) becomes

$$\begin{aligned}
&-\mathbb{E}[\langle\mathcal{L}_{ll'}^2\rangle - \langle\mathcal{L}_{ll'}\rangle^2] \\
&\quad= \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[W_{il}^*\langle w_{il'}w_{jl}w_{jl'}\rangle - W_{il}^*W_{jl'}^*\langle w_{il'}w_{jl}\rangle - W_{il}^*W_{jl}^*\langle w_{il'}w_{jl'}\rangle \\
&\qquad - \big\{W_{il}^*\langle w_{il'}w_{jl'}w_{jl}\rangle - W_{il}^*\langle w_{il'}w_{jl}\rangle\langle w_{jl'}\rangle\big\} \\
&\qquad - W_{il}^*\langle w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle + W_{il}^*W_{jl'}^*\langle w_{il'}\rangle\langle w_{jl}\rangle + W_{il}^*W_{jl}^*\langle w_{il'}\rangle\langle w_{jl'}\rangle \\
&\qquad + \big\{W_{il}^*\langle w_{il'}\rangle\langle w_{jl'}w_{jl}\rangle - 2W_{il}^*\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + W_{il}^*\langle w_{il'}w_{jl'}\rangle\langle w_{jl}\rangle\big\}\Big] + \mathcal{O}(K^2/n) \\
&\quad\overset{\text{N}}{=} \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[\langle w_{il}\rangle\langle w_{il'}w_{jl}w_{jl'}\rangle - \langle w_{il}w_{jl'}\rangle\langle w_{il'}w_{jl}\rangle - \langle w_{il}w_{jl}\rangle\langle w_{il'}w_{jl'}\rangle - \langle w_{il}\rangle\langle w_{il'}w_{jl'}w_{jl}\rangle \\
&\qquad + \langle w_{il}\rangle\langle w_{il'}w_{jl}\rangle\langle w_{jl'}\rangle - \langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle + \langle w_{il}w_{jl'}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle + \langle w_{il}w_{jl}\rangle\langle w_{il'}\rangle\langle w_{jl'}\rangle \\
&\qquad + \langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl'}w_{jl}\rangle - 2\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{il}\rangle\langle w_{il'}w_{jl'}\rangle\langle w_{jl}\rangle\Big] + \mathcal{O}(K^2/n) \\
&\quad= \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[- \langle w_{il}w_{jl'}\rangle\langle w_{il'}w_{jl}\rangle - \langle w_{il}w_{jl}\rangle\langle w_{il'}w_{jl'}\rangle + 2\langle w_{il}w_{jl'}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle \\
&\qquad + \langle w_{il}w_{jl}\rangle\langle w_{il'}\rangle\langle w_{jl'}\rangle + \langle w_{il'}w_{jl'}\rangle\langle w_{il}\rangle\langle w_{jl}\rangle - 2\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle\Big] + \mathcal{O}(K^2/n) \\
&\quad= -\frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[\Big(\langle w_{il'}w_{jl}\rangle - \langle w_{il'}\rangle\langle w_{jl}\rangle\Big)\Big(\langle w_{il}w_{jl'}\rangle - \langle w_{il}\rangle\langle w_{jl'}\rangle\Big) \\
&\qquad + \Big(\langle w_{il}w_{jl}\rangle - \langle w_{il}\rangle\langle w_{jl}\rangle\Big)\Big(\langle w_{il'}w_{jl'}\rangle - \langle w_{il'}\rangle\langle w_{jl'}\rangle\Big)\Big] + \mathcal{O}(K^2/n)
\end{aligned} \tag{97}$$

which is (74) once expressed with the overlaps (90), (91). ■

B.3 Derivation of (75)

Let us now compute the following term:

$$\begin{aligned}\mathbb{E}[\langle\mathcal{L}_{ll'}\rangle^2] = \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[&\langle w_{il}w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle + \langle w_{il}\rangle\langle w_{jl}\rangle W^*_{il'}W^*_{jl'} + \langle w_{il'}\rangle\langle w_{jl'}\rangle W^*_{il}W^*_{jl}\\ &- 2\langle w_{il}w_{il'}\rangle\langle w_{jl}\rangle W^*_{jl'} - 2\langle w_{il}w_{il'}\rangle\langle w_{jl'}\rangle W^*_{jl} + 2\langle w_{il}\rangle\langle w_{jl'}\rangle W^*_{il'}W^*_{jl}\\ &+ 4\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big(\langle \underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_j\Big) - 4\langle w_{jl}w_{jl'}\rangle\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\\ &+ 4\langle w_{jl}\rangle W^*_{jl'}\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big) + 4\langle w_{jl'}\rangle W^*_{jl}\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big].\end{aligned} \tag{98}$$

Now we need to integrate by parts all the noise dependent terms (again, it is necessary that this operation is done before using the Nishimori identity):

$$\begin{aligned}&-4\mathbb{E}\Big[\langle w_{jl}w_{jl'}\rangle\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big]\\ &\qquad\overset{(88)}{=} -2\mathbb{E}\Big[\langle w_{jl}w_{jl'}\rangle\langle w_{il}w_{il'}\rangle - 2\langle w_{jl}w_{jl'}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + \langle w_{il}\rangle\langle w_{jl}w_{jl'}w_{il'}\rangle\Big], \qquad (99)\\ &4\mathbb{E}\Big[\langle w_{jl}\rangle W^*_{jl'}\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big]\\ &\qquad\overset{(88)}{=} 2\mathbb{E}\Big[W^*_{jl'}\langle w_{jl}\rangle\langle w_{il}w_{il'}\rangle - 2W^*_{jl'}\langle w_{jl}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + W^*_{jl'}\langle w_{il}\rangle\langle w_{jl}w_{il'}\rangle\Big], \qquad (100)\\ &4\mathbb{E}\Big[\langle w_{jl'}\rangle W^*_{jl}\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big]\\ &\qquad\overset{(88)}{=} 2\mathbb{E}\Big[W^*_{jl}\langle w_{jl'}\rangle\langle w_{il}w_{il'}\rangle - 2W^*_{jl}\langle w_{jl'}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + W^*_{jl}\langle w_{il}\rangle\langle w_{jl'}w_{il'}\rangle\Big]. \qquad (101)\end{aligned}$$

We now tackle the more painful term:

$$\begin{aligned}&4\mathbb{E}\Big[\Big(\langle \underline{w}_i^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_i\Big)\Big(\langle \underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_j\Big)\Big]\\ &\quad\overset{(88)}{=} 2\mathbb{E}\Big[\Big(\langle \underline{w}_j^\intercal\rangle\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_j\Big)\big(\langle w_{il}w_{il'}\rangle - 2\langle w_{il}\rangle\langle w_{il'}\rangle\big)\\ &\qquad + \langle w_{il}\rangle\Big\langle\Big(\underline{w}_j^\intercal\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\underline{\widehat{Z}}_j\Big)w_{il'}\Big\rangle + 2\sum_{k,k'=1}^{K}\sum_{u=1}^{K}\langle w_{ik}\rangle\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{kk'}\langle w_{ju}\rangle\Big(\frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\Big)_{uk'}\delta_{ji}\Big]\end{aligned} \tag{102}$$

where the last term comes from the contribution when the two noise variables are equal (i.e. the contribution corresponding to the potential explicit dependence of A in the noise $\underline{\widehat{Z}}_i$ in (88)). Note that this last term (the term with a δ_{ji}) is $\mathcal{O}(K^3/n)$. Now we again integrate by part w.r.t. the second noise variable the r.h.s. of this

last identity term after term:

$$
\begin{aligned}
&2\mathbb{E}\Big[\Big(\langle \underline{w}_j^\intercal\rangle \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big)\langle w_{il}w_{il'}\rangle\Big] \\
&\qquad \overset{(88)}{=} \mathbb{E}\Big[\langle w_{il}w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle - 2\langle w_{il}w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{jl}\rangle\langle w_{il}w_{il'}w_{jl'}\rangle\Big], \qquad (103)\\
&-4\mathbb{E}\Big[\Big(\langle \underline{w}_j^\intercal\rangle \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big)\langle w_{il}\rangle\langle w_{il'}\rangle\Big] \\
&\qquad \overset{(87)}{=} -2\mathbb{E}\Big[\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle - 3\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{jl}\rangle\langle w_{il'}\rangle\langle w_{il}w_{jl'}\rangle \\
&\qquad\qquad + \langle w_{jl}\rangle\langle w_{il}\rangle\langle w_{il'}w_{jl'}\rangle\Big], \qquad (104)\\
&2\mathbb{E}\Big[\langle w_{il}\rangle\Big\langle\Big(\underline{w}_j^\intercal \frac{d\epsilon^{1/2}}{d\epsilon_{ll'}}\widehat{\underline{Z}}_j\Big)w_{il'}\Big\rangle\Big] \\
&\qquad \overset{(89)}{=} \mathbb{E}\Big[\langle w_{il}\rangle\langle w_{il'}w_{jl}w_{jl'}\rangle - 2\langle w_{il}\rangle\langle w_{il'}w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{il'}w_{jl}\rangle\langle w_{il}w_{jl'}\rangle\Big]. \qquad (105)
\end{aligned}
$$

We are now ready to combine all terms in (98). Using the Nishimori identity it yields

$$
\begin{aligned}
\mathbb{E}[\langle\mathcal{L}_{ll'}\rangle^2] &\overset{\text{N}}{=} \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[\langle w_{il}w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle + \langle w_{il}\rangle\langle w_{jl}\rangle\langle w_{il'}w_{jl'}\rangle + \langle w_{il'}\rangle\langle w_{jl'}\rangle\langle w_{il}w_{jl}\rangle \\
&- 2\langle w_{il}w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle - 2\langle w_{il}w_{il'}\rangle\langle w_{jl'}\rangle\langle w_{jl}\rangle + 2\langle w_{il}\rangle\langle w_{jl'}\rangle\langle w_{il'}w_{jl}\rangle \\
&- 2\Big\{\langle w_{jl}w_{jl'}\rangle\langle w_{il}w_{il'}\rangle - 2\langle w_{jl}w_{jl'}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + \langle w_{il}\rangle\langle w_{jl}w_{jl'}w_{il'}\rangle\Big\} \\
&+ 2\Big\{\langle w_{jl'}\rangle\langle w_{jl}\rangle\langle w_{il}w_{il'}\rangle - 2\langle w_{jl'}\rangle\langle w_{jl}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + \langle w_{jl'}\rangle\langle w_{il}\rangle\langle w_{jl}w_{il'}\rangle\Big\} \\
&+ 2\Big\{\langle w_{jl}\rangle\langle w_{jl'}\rangle\langle w_{il}w_{il'}\rangle - 2\langle w_{jl}\rangle\langle w_{jl'}\rangle\langle w_{il}\rangle\langle w_{il'}\rangle + \langle w_{jl}\rangle\langle w_{il}\rangle\langle w_{jl'}w_{il'}\rangle\Big\} \\
&+ \begin{cases}
\langle w_{il}w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle - 2\langle w_{il}w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{jl}\rangle\langle w_{il}w_{il'}w_{jl'}\rangle \\
-2\big\{\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}w_{jl'}\rangle - 3\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{jl}\rangle\langle w_{il'}\rangle\langle w_{il}w_{jl'}\rangle + \langle w_{jl}\rangle\langle w_{il}\rangle\langle w_{il'}w_{jl'}\rangle\big\} \\
\langle w_{il}\rangle\langle w_{il'}w_{jl}w_{jl'}\rangle - 2\langle w_{il}\rangle\langle w_{il'}w_{jl}\rangle\langle w_{jl'}\rangle + \langle w_{il'}w_{jl}\rangle\langle w_{il}w_{jl'}\rangle
\end{cases} \\
&\qquad\qquad \Big] + \mathcal{O}(K^3/n) \\
&= \frac{1}{4n^2}\sum_{i,j=1}^{n}\mathbb{E}\Big[\langle w_{il}\rangle\langle w_{jl}\rangle\langle w_{il'}w_{jl'}\rangle + \langle w_{il'}\rangle\langle w_{jl'}\rangle\langle w_{il}w_{jl}\rangle \\
&\qquad\qquad + \langle w_{il'}w_{jl}\rangle\langle w_{il}w_{jl'}\rangle - 2\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle\Big] + \mathcal{O}(K^3/n) \\
&= \frac{1}{4}\mathbb{E}\Big[\langle Q_{l'l}\rangle^2 + \langle Q_{ll'}\rangle^2 + \langle Q_{ll'}Q_{l'l}\rangle - \frac{2}{n^2}\sum_{i,j=1}^{n}\langle w_{il}\rangle\langle w_{il'}\rangle\langle w_{jl}\rangle\langle w_{jl'}\rangle\Big] + \mathcal{O}(K^3/n) \qquad (106)
\end{aligned}
$$

using (90), (91) for the last identity. Combining this with (85) (once squarred) we finally obtain (75). ■

B.4 Derivation of (77)

First note from (90), (91) that the "thermal fluctuation" of the overlaps can be written as

$$\mathbb{E}\big\langle (Q_{ll'} - \langle Q_{ll'}\rangle)^2 \big\rangle = \mathbb{E}\langle Q_{ll'}^2\rangle - \mathbb{E}[\langle Q_{ll'}\rangle^2] = \mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n}\langle w_{il}w_{jl}\rangle\big(\langle w_{il'}w_{jl'}\rangle - \langle w_{il'}\rangle\langle w_{jl'}\rangle\big)$$
$$\leq \Big\{\mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n}\langle w_{il}w_{jl}\rangle^2\Big\}^{1/2}\Big\{\mathbb{E}\frac{1}{n^2}\sum_{i,j=1}^{n}\big(\langle w_{il'}w_{jl'}\rangle - \langle w_{il'}\rangle\langle w_{jl'}\rangle\big)^2\Big\}^{1/2} \tag{107}$$

using Cauchy-Schwarz. Note that the first term of the r.h.s. of this last identity is bounded. Now the formula (97) for the special case $l = l'$ yields

$$\mathbb{E}\big\langle (\mathcal{L}_{l'l'} - \langle \mathcal{L}_{l'l'}\rangle)\big\rangle^2 = \mathbb{E}\frac{1}{2n^2}\sum_{i,j=1}^{n}\big(\langle w_{il'}w_{jl'}\rangle - \langle w_{il'}\rangle\langle w_{jl'}\rangle\big)^2 + \mathcal{O}(K^2/n). \tag{108}$$

Thus we obtain (77).

C 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 [41, 35, 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 $f : \mathbb{R}^n \to \mathbb{R}$:

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\log f = \lim_{p\to 0^+}\lim_{n\to\infty}\frac{1}{np}\log \mathbb{E} f^p. \tag{109}$$

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 (109) 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:

$$\mathbb{E}Z_n^p = \mathbb{E}\left[\int_{\mathbb{R}^n\times\mathbb{R}^K} \mathrm{d}w \prod_{i=1}^{n} P_0\left(\{w_{il}\}_{l=1}^{K}\right)\prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu\middle|\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_{\mu i}w_{il}\right\}_{l=1}^{K}\right)\right]^p,$$
$$= \mathbb{E}\left[\prod_{a=1}^{p}\int_{\mathbb{R}^n\times\mathbb{R}^K} \mathrm{d}w^a \prod_{i=1}^{n} P_0\left(\{w_{il}^a\}_{l=1}^{K}\right)\prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu\middle|\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_{\mu i}w_{il}^a\right\}_{l=1}^{K}\right)\right].$$

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

$$\mathbb{E}Z_n^p = \mathbb{E}_X \int_{\mathbb{R}^m} \mathrm{d}Y \prod_{a=0}^{p} \left[\int_{\mathbb{R}^n \times \mathbb{R}^K} \mathrm{d}w^a \prod_{i=1}^{n} P_0\left(\{ w_{il}^a \}_{l=1}^K\right) \right.$$
$$\left. \times \prod_{\mu=1}^{m} P_{\text{out}}\left(Y_\mu \middle| \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{\mu i} w_{il}^a \right\}_{l=1}^K \right) \right].$$

To perform the average over X, we notice that, since it is an iid 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:

$$\mathbb{E}Z_{\mu l}^a Z_{\nu l'}^b = \delta_{\mu\nu} \Sigma_{\substack{al \\ bl'}} = \delta_{\mu\nu} Q_{bl'}^{al}, \tag{110}$$

$$Q_{bl'}^{al} \equiv \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b. \tag{111}$$

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 :

$$\mathbb{E}\left[Z(Y)^n\right] = \prod_{(a,r)} \int_{\mathbb{R}} \mathrm{d}Q_{ar}^{ar} \prod_{\{(a,r);(b,r')\}} \int_{\mathbb{R}} \mathrm{d}Q_{br'}^{ar} \left[I_{\text{prior}}(\{Q_{br'}^{ar}\}) \times I_{\text{channel}}(\{Q_{br'}^{ar}\}) \right], \tag{112}$$

with:

$$I_{\text{prior}}(\{Q_{br'}^{ar}\}) = \prod_{a=0}^{p} \left[\int_{\mathbb{R}^{n \times K}} \mathrm{d}w^a P_0(w^a) \right] \left[\prod_{\{(a,l);(b,l')\}} \delta\left(Q_{bl'}^{al} - \frac{1}{n} \sum_{i=1}^{n} w_{il}^a w_{il'}^b \right) \right], \tag{113}$$

$$I_{\text{channel}}(\{Q_{br'}^{ar}\}) = \int_{\mathbb{R}^m} \mathrm{d}Y \prod_{a=0}^{p} \int_{\mathbb{R}^{m \times K}} \mathrm{d}Z^a \prod_{a=0}^{p} P_{\text{out}}(Y|Z^a) e^{-\frac{m}{2} \log\det\Sigma - \frac{mK(p+1)}{2} \log 2\pi}$$
$$\exp\left[-\frac{1}{2} \sum_{\mu=1}^{m} \sum_{a,b} \sum_{l,l'} Z_{\mu l}^a Z_{\mu l'}^b (\Sigma^{-1})_{\substack{al \\ bl'}} \right]. \tag{114}$$

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

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[Z(Y)^p\right] = \text{extr}_{Q,\hat{Q}} \left[H(Q, \hat{Q}) \right], \tag{115}$$

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

$$H(Q,\hat{Q}) \equiv \frac{1}{2} \sum_{a=0}^{p} \sum_{l,l'} Q_{al}^{al} \hat{Q}_{al}^{al} - \frac{1}{2} \sum_{a \neq b} \sum_{l,l'} Q_{bl'}^{al} \hat{Q}_{bl'}^{al} + \log I + \alpha \log J, \tag{116}$$

in which we defined:

$$I \equiv \prod_{a=0}^{p} \int_{\mathbb{R}^K} \mathrm{d}w^a P_0(w^a) \exp\left[-\frac{1}{2}\sum_{a=0}^{p}\sum_{l,l'} \hat{Q}^{al}_{al'} w^a_l w^a_{l'} + \frac{1}{2}\sum_{a\neq b}\sum_{l,l'} \hat{Q}^{al}_{bl'} w^a_l w^b_{l'}\right], \tag{117}$$

$$J \equiv \int_{\mathbb{R}} \mathrm{d}y \prod_{a=0}^{p} \int_{\mathbb{R}^K} \frac{\mathrm{d}Z^a}{(2\pi)^{K(p+1)/2}} \frac{P_{out}(y|Z^a)}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}\sum_{a,b=0}^{p}\sum_{l,l'=1}^{K} Z^a_l Z^b_{l'} (\Sigma^{-1})^{al}_{bl'}\right]. \tag{118}$$

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:

$$\exists Q^0 \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall a \in [|0,p|] \quad \forall (l,l') \in [|1,K|]^2 \quad Q^{al}_{al'} = Q^0_{ll'}, \tag{119}$$

$$\exists q \in \mathbb{R}^{K\times K} \text{ s.t} \quad \forall (a<b) \in [|0,p|]^2 \quad \forall (l,l') \in [|1,K|]^2 \quad Q^{al}_{bl'} = q_{ll'}, \tag{120}$$

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:

$$H(Q^0,\hat{Q}^0,q,\hat{q}) = \frac{p+1}{2}\mathrm{Tr}[Q^0\hat{Q}^0] - \frac{p(p+1)}{2}\mathrm{Tr}[q\hat{q}] + \log I + \alpha \log J. \tag{121}$$

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:

$$I = \int_{\mathbb{R}^K} \mathcal{D}\xi \left[\int_{\mathbb{R}^K} \mathrm{d}w \, P_0(w) \, \exp\left[-\frac{1}{2} w^\intercal (\hat{Q}^0 + \hat{q}) w + \xi^\intercal \hat{q}^{1/2} w\right]\right]^{p+1}, \tag{122}$$

$$J = \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathcal{D}\xi \left[\int_{\mathbb{R}^K} \mathrm{d}Z P_{\mathrm{out}}\left\{y|(Q^0-q)^{1/2}Z + q^{1/2}\xi\right\}\right]^{p+1}. \tag{123}$$

Our assumptions must be consistent in the sense that $\mathrm{extr}_{Q,\hat{Q}}\left[\lim_{p\to 0^+} H(Q,\hat{Q})\right] = 0$ (because $\mathbb{E}Z_n^0 = 1$). In the $p \to 0^+$ limit, one easily gets $J = 1$ and $I = \int_{\mathbb{R}^K} \mathrm{d}w \, P_0(w) \exp\left[-\frac{1}{2} w^\intercal \hat{Q}^0 w^0\right]$. 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:

$$f = \mathrm{extr}_{q,\hat{q}} \left\{-\frac{1}{2}\mathrm{Tr}[q\hat{q}] + I_P + \alpha I_C\right\}, \tag{124}$$

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

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

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.

D 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:

$$\epsilon_g^{\text{Gibbs}} \equiv \frac{1}{2}\mathbb{E}_{W^*,X}\big\langle \left[\varphi_{\text{out}}\left(XW\right) - \varphi_{\text{out}}\left(XW^*\right)\right]^2 \big\rangle, \tag{125}$$

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

$$\epsilon_g^{\text{Bayes}} \equiv \frac{1}{2}\mathbb{E}_{W^*,X}\left[\left(\left\langle\varphi_{\text{out}}\left(XW\right)\right\rangle - \varphi_{\text{out}}\left(XW^*\right)\right)^2\right]. \tag{126}$$

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 A.4, 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 [42]), the generalization error can itself be evaluated as a function of q^*. Example where is done includes e.g. [43, 3, 19, 11].

E The large K limit in the committee symmetric setting

We consider the large K limit[3] 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 matrices q and $\hat{q}$ (q and R in the notations of Theorem 3.1) are of the type $q = q_d I_{K\times K} + \frac{q_a}{K}\mathbf{1}_K\mathbf{1}_K^\intercal$, 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 $Q^0 = I_{K\times K}$ ($Q^0 = \rho$ 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

[3] A similar limit has been derived in the context of coding with sparse superposition codes [44]. There the large input alphabet limit of the mutual information is considered *after* the thermodynamc limit $n \to \infty$ corresponding to the large codeword limit in this coding context.

$0 \leq q_a + q_d \leq 1$.

E.1 Large K limit for sign activation function

In the following, we consider $Q^0 = \sigma^2 I_{K\times K}$. We are interested here in computing the leading order term in I_C of (124). Note that replacing σ^2 by 1 in this equation only amounts to replacing q by q/σ^2, so we can assume $\rho = 1$ without loss of generality. We write I_C in (124) as $I_C = \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathcal{D}\xi\, I_C(y,\xi) \log I_C(y,\xi)$. A simple theoretical physics calculation yields the expression:

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

Denote

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

and for $1 \leq l \leq R$, one can rewrite the factorized integral in the last expression of I_C as:

$$J_l \equiv e^{-\frac{\lambda_l^2}{2}+i\lambda_l\frac{\hat{w}}{\sqrt{K}}} \int_{\mathbb{R}} \mathcal{D}z e^{z(\lambda_l - i\frac{\hat{w}}{\sqrt{K}})} e^{-\frac{i\hat{u}}{\sqrt{K}}\mathrm{sign}[z]},$$

and $J \equiv \prod_{l=1}^{K} J_l$. Using that:

$$F(\alpha, i\beta) \equiv \int_{\mathbb{R}} \mathcal{D}z e^{\alpha z + i\beta\,\mathrm{sign}(z)} = e^{\alpha^2/2}\left[\cos\beta + i\sin\beta \hat{H}(\alpha)\right], \tag{127}$$

with $\hat{H}(x) = \mathrm{erf}(x/\sqrt{2})$, we obtain:

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

Note that we have $\lambda_l = \lambda_{l,0} + \frac{1}{\sqrt{K}}\lambda_1$ with:

$$\lambda_{l,0} = \sqrt{\frac{q_d}{1-q_d}}\xi_l \quad \text{and} \quad \lambda_1 = \left[\sqrt{\frac{1-q_a-q_d}{1-q_d}}-1\right]w + \left[\sqrt{\frac{q_a+q_d}{1-q_d}} - \sqrt{\frac{q_d}{1-q_d}}\right]\frac{\mathbf{1}_K^\intercal \xi}{\sqrt{K}}.$$

Expanding J_l as $K \to \infty$, we obtain:

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

Then we have $J = \exp\left[\sum_{l=1}^{K} \ln J_l\right]$, which yields:

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

in which we defined

$$w_\xi \equiv \frac{1}{\sqrt{K}}\sum_{l=1}^{K}\xi_l, \qquad \Gamma_0 \equiv \frac{1}{K}\sum_{l=1}^{K} e^{-\frac{1}{2}\lambda_{l,0}^2}, \qquad S_1 \equiv \frac{1}{\sqrt{K}}\sum_{l=1}^{K}\hat{H}(\lambda_{l,0}), \qquad S_2 \equiv \frac{1}{K}\sum_{l=1}^{K}\hat{H}(\lambda_{l,0})^2.$$

A detailed calculation actually shows that the calculation of J is true not only up to $\mathcal{O}(K^{-1/2})$ but to $\mathcal{O}(K^{-1})$. Recall that $I(y,\boldsymbol{\xi}) = (4\pi^2)^{-1}\int \mathrm{d}w\mathrm{d}\hat{w}\mathrm{d}u\mathrm{d}\hat{u}e^{iw\hat{w}+iu\hat{u}}\delta_{y,\mathrm{sign}(u)} \times J$. One can now readily perform the integration over all variables to obtain:

$$I(y,\boldsymbol{\xi}) = H\left[-y\frac{S_1 + \sqrt{\frac{2}{\pi}}w_\xi\Gamma_0\frac{\sqrt{q_d+q_a}-\sqrt{q_d}}{\sqrt{1-q_d}}}{\sqrt{1-S_2-\frac{2}{\pi}\Gamma_0^2\frac{q_a}{1-q_d}}}\right] + \mathcal{O}(K^{-1}), \tag{128}$$

in which $H(x) \equiv \int_x^\infty \mathcal{D}z = \frac{1}{2}(1-\mathrm{erf}(x/\sqrt{2}))$. Note that all quantities $w_\xi, \Gamma_0, S_1, S_2$ only depend on ξ via its empirical measure, which means the integration over $\xi \in \mathbb{R}^K$ is tractable. We compute it in the following, using theoretical physics methods. Basically, denoting

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

it amounts to write:

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

Computing the last integral when $K\to\infty$, and defining $\gamma \equiv \frac{2}{\pi}(q_a + \arcsin q_d)$, one reduces this form to the final expression:

$$\begin{aligned}
I_C &= \sum_{y=\pm1}\int_{\mathbb{R}}\mathcal{D}xH\left[yx\sqrt{\frac{\gamma}{1-\gamma}}\right]\log H\left[yx\sqrt{\frac{\gamma}{1-\gamma}}\right] + \mathcal{O}(K^{-1}),\\
I_C &= 2\int_{\mathbb{R}}\mathcal{D}xH\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right]\log H\left[x\sqrt{\frac{\gamma}{1-\gamma}}\right] + \mathcal{O}(K^{-1}).
\end{aligned} \tag{129}$$

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

E.2 The Gaussian prior

The prior part I_P of the free entropy (124) is very easy to evaluate in the Gaussian prior setting. We consider a prior with variance $\rho = 1$ (we can simply rescale q by q/ρ in the final expression for a finite variance ρ). We obtain:

$$I_P = \frac{K}{2}\hat{q}_d + \frac{1}{2}\hat{q}_a - \frac{K-1}{2}\log(1+\hat{q}_d) - \frac{1}{2}\log\left(1+\hat{q}_d+\hat{q}_a\right). \tag{130}$$

E.3 The fixed point equations

From the free entropy (124) and (129), (130), one easily obtains the fixed point equations (after having extremized over $\hat{q}_d$ and $\hat{q}_a$) as (recall $\alpha = \lim \frac{m}{n}$):

$$\partial_{q_a}(I_G + \alpha I_C) = 0, \tag{131}$$
$$\partial_{q_d}(I_G + \alpha I_C) = 0, \tag{132}$$

with

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

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

The fixed point equations (131), (132) have different behaviors depending on the scaling of α with the hidden layer size K. We detail these different behaviors in the following.

E.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 can be simplified as:

$$\begin{cases} q_d = 0, \\ q_a = 2\alpha(1-q_a)\frac{\partial \mathcal{I}_C}{\partial q_a}. \end{cases} \tag{133}$$

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

In this regime, we naturally define $\alpha = \widetilde{\alpha}K$, with $\widetilde{\alpha}$ of order 1. One can show that the fixed point equations (131), (132) only admit a solution with the following scaling : $q_a + q_d = 1 - \frac{\chi}{K}$. The fixed point equations in terms of χ and q_d read:

$$\begin{cases} \gamma = \frac{2}{\pi}\left(\arcsin(q_d) - q_d + 1 - \frac{\chi}{K}\right) = \frac{2}{\pi}\left(\arcsin(q_d) - q_d + 1\right) + \mathcal{O}(K^{-1}), \\ q_d = 2(1-q_d)\left(\frac{1}{\sqrt{1-q_d^2}} - 1\right)\widetilde{\alpha}\frac{\partial \mathcal{I}_C}{\partial q_a}, \\ \chi^{-1} = 2\widetilde{\alpha}\frac{\partial \mathcal{I}_C}{\partial q_a}. \end{cases} \tag{134}$$

The State Evolution (SE) computation of Figure 1 was performed by solving the fixed point equations (133), (134) (depending on the regime of α).

The stability of the $q_d = 0$ solution: It is easy to show that (134) always admit a non-specialized solution with $q_d = 0$. This solution stops however to be optimal in terms of the free energy at a finite value of $\widetilde{\alpha}_{\text{spec}} \simeq 7.65$. However, this solution will remain linearly stable for every $\widetilde{\alpha}$. Actually, one can show that this non-specialized solution will remain linearly stable for α up to order $\Theta(K^2)$. Indeed, adding the correct time

indices to iterate the state evolution fixed point equations, we obtain:

$$q_d^{t+1} = \frac{F(q_d^t, q_a^t)}{1 + F(q_d^t, q_a^t)}, \tag{135}$$

$$q_a^{t+1} = \frac{G(q_d^t, q_a^t)}{\left(1 + F(q_d^t, q_a^t)\right)\left(1 + F(q_d^t, q_a^t) G(q_d^t, q_a^t)\right)}, \tag{136}$$

with F and G defined as:

$$F(q_d^t, q_a^t) \equiv \frac{2\alpha}{K-1}\left[\partial_{q_d} I_C - \partial_{q_a} I_C\right], \tag{137}$$

$$G(q_d^t, q_a^t) \equiv \frac{2\alpha K}{K-1}\left[\partial_{q_a} I_C - \frac{1}{K}\partial_{q_d} I_C\right]. \tag{138}$$

We focusing on the behavior of (135) around $q_d = 0$. Since we have shown that in the $K \to \infty$ limit, the leading order in I_C only depends on $\gamma = \frac{1}{\pi}(q_a + \arcsin q_d)$, one easily computes that for $\alpha = o(K^2)$ (the regime in which only the leading order of I_C contributes), $\frac{\partial F}{\partial q_d}|_{q_d=0} \to_{K\to\infty} 0$, 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. E.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, see (137). To summarize, the non-specialized solution always remains linearly stable in the large K limit for α of order smaller than K^2. This implies that Approximate Message Passing, implemented in such a regime, could not possibly find the specialized solution in this regime. Note that this range of scaling of α is possibly broader, as one would have to explicitly compute $I_C^{(1)}$ in order to check that $\frac{\partial F}{\partial q_d}|_{q_d=0} \neq 0$. This tedious calculation is left for future work.

E.4 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. D), one can directly give an explicit expression of this error in the $K = 2$ case. Recall that the overlap matrix $q = q_d I_{K\times K} + \frac{q_a}{K}\mathbf{1}_K\mathbf{1}_K^\intercal$ with $(\mathbf{1}_K)_l = 1$. For simplicity, we denote $\text{sign}(x) = \sigma(x)$. One obtains:

$$\begin{aligned}\frac{1}{2} - 2\epsilon_g^{\text{Bayes,K=2}} = &\int_{\mathbb{R}^4} \mathcal{D}x\, \sigma\left[\sigma(x_1) + \sigma(x_2)\right] \\ &\times \sigma\left\{\sigma\left[(\frac{q_a}{2} + q_d)x_1 + \frac{q_a}{2}x_2 + x_3\sqrt{1 - \frac{q_a^2}{2} - q_a q_d - q_d^2}\right]\right. \\ &\left. + \sigma\left[\frac{q_a}{2}x_1 + (\frac{q_a}{2} + q_d)x_2 - x_3\frac{q_a(q_d + \frac{q_a}{2})}{\sqrt{1 - \frac{q_a^2}{2} - q_a q_d - q_d^2}} + x_4\sqrt{\frac{(1-q_d^2)(1-(q_a+q_d)^2)}{1 - \frac{q_a^2}{2} - q_a q_d - q_d^2}}\right]\right\}.\end{aligned} \tag{139}$$

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. 1.

E.5 The generalization error at large K

Recall the definition of the generalization error from sec. D. From the remarks of the section, one can compute it using (125), by noting that $\epsilon_g^{\text{Gibbs}} = \epsilon_g(q^*)$, in which q^* is the optimal overlap, and where we defined:

$$\epsilon_g(q) \equiv \frac{1}{2}\mathbb{E}_X \left[\varphi_{\text{out}}\left(XW\right) - \varphi_{\text{out}}\left(XW^*\right)\right]^2 .$$

This quantity indeed only depends on the overlap $q = W^\intercal W^*$ by rotation invariance of the distribution of X. At large K, one can apply the same expansion used for computing I_C in sec. E.1, and obtains after a tedious yet straightforward calculation:

$$\epsilon_g^{\text{Bayes}} = \frac{1}{2}\epsilon_g^{\text{Gibbs}} = \frac{1}{\pi}\arccos\left[\frac{2}{\pi}\left(q_a + \arcsin q_d\right)\right] + \mathcal{O}(K^{-1}). \tag{140}$$

This expression is the one used in the computation of Fig. 1.

F 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 $\varphi_{\text{out}}(\mathbf{h}) = \sigma\left(\frac{1}{\sqrt{K}}\sum_{l=1}^K h_l\right)$. 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 (124). 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 (124) gives I_C as $I_C = \int_{\mathbb{R}} \mathrm{d}y \int_{\mathbb{R}^K} \mathcal{D}\xi I_C(y,\xi)\log I_C(y,\xi)$. One can easily derive:

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

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

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

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

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

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

$$\int_{\mathbb{R}^K}\mathcal{D}\xi F(x^\intercal\xi) = \frac{1}{\sqrt{2\pi x^\intercal x}}\int_{\mathbb{R}}\mathrm{d}u F(u)e^{-\frac{u^2}{2x^\intercal x}}. \tag{141}$$

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

$$I_C = \int_{\mathbb{R}} \mathrm{d}y \frac{1}{\sqrt{2\pi\Gamma(q)}} \int_{\mathbb{R}} \mathrm{d}v e^{-\frac{v^2}{2\Gamma(q)}} I_C(v,y) \log I_C(v,y), \tag{142}$$

$$I_C(v,y) \equiv \frac{1}{\sqrt{2\pi(1-\Gamma(q))}} \int_{\mathbb{R}} \mathrm{d}u \, P_{\text{out}}(y|\sigma(u)) \exp\left[-\frac{1}{2\left(1-\Gamma(q)\right)}(u-v)^2\right]. \tag{143}$$

Note that by hypothesis, both q and $\mathbb{1} - q$ 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.

G Update functions and AMP derivation

AMP can be seen as Taylor expansion of the Loopy Belief Propagation approach [13, 14, 45], similar to the so-called Thouless-Anderson-Palmer equation in spin glass theory [34]. While the behaviour of AMP can be rigorously studied [17, 18, 46], it is useful and instructive to see how the derivation can be performed in the framework of Belief Propagation (BP) and the cavity method, as was pioneered in [35, 37] for the single layer problem. The derivation uses the Generalized AMP notations of [16] and follows closely the one of [26].

G.1 Definition of the update functions

Let's consider the distributions probabilities $\tilde{P}_{\text{out}}$ and $\tilde{P}_0$, related up to a normalizing constant, to the problems 4 and 7. 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, as the mean and variance with respect to these distributions:

$$\begin{cases} \tilde{P}_{\text{out}}(z;\omega,y,V) \equiv \frac{1}{\mathcal{Z}_{\text{out}}} e^{-\frac{1}{2}(z-\omega)^\intercal V^{-1}(z-\omega)} P_{\text{out}}(y|z) \\ g_{\text{out}}(\omega,y,V) = \frac{1}{\mathcal{Z}_{\text{out}}} \frac{\partial \mathcal{Z}_{\text{out}}}{\partial \omega}(\omega,y,V) = V^{-1}\mathbb{E}_{\tilde{P}_{\text{out}}}[z-\omega] \\ \partial_\omega g_{\text{out}}(\omega,y,V) = V^{-1}\mathbb{E}_{\tilde{P}_{\text{out}}}[(z-\omega)(z-\omega)^\intercal] - V^{-1} - g_{\text{out}} g_{\text{out}}^\intercal \end{cases}$$

$$\begin{cases} \tilde{P}_0(w;\Sigma,T) \equiv \frac{1}{\mathcal{Z}_{P_0}} P_0(w) e^{-\frac{1}{2}(w-T)^\intercal \Sigma^{-1}(w-T)} \\ f_W(\Sigma,T) = \mathbb{E}_{\tilde{P}_0}[w] \\ f_C(\Sigma,T) = \mathbb{E}_{\tilde{P}_0}[ww^\intercal] - f_W f_W^\intercal \end{cases}$$

G.2 Approximate message passing algorithm

G.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. Beliefs propagation equations can be formulated as the following [14, 45], where $w_i = (w_{il})_{l=1..K} \in \mathbb{R}^K$.

$$\begin{cases} m_{i\to\mu}(t+1,w_i) = \frac{1}{\mathcal{Z}_{i\to\mu}} P_0(w_i) \prod_{k\neq\mu}^{m} \tilde{m}_{\nu\to i}(t,w_i) \\ \tilde{m}_{\mu\to i}(t,w_i) = \frac{1}{\mathcal{Z}_{\mu\to i}} \int \prod_{j\neq i}^{n} dw_j P_{\text{out}}\left(Y_\mu | \sum_{j=1}^{n} X_{\mu j} w_j\right) m_{j\to\mu}(t,w_j) \end{cases} \tag{144}$$

where we absorbed the factor $\frac{1}{\sqrt{n}}$ in the element $X_{\mu i}$, which are therefor of order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. The term inside P_{out} can be decouple using its K-dimensional Fourrier transform:

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

From this, (144) becomes:

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

and we define the mean and variance of the messages:

$$\begin{cases}\hat{W}_{j\to\mu}(t) \equiv \int_{\mathbb{R}^K} dw_j m_{j\to\mu}(t,w_j)w_j\\ \hat{C}_{j\to\mu}(t) \equiv \int_{\mathbb{R}^K} dw_j m_{j\to\mu}(t,w_j)w_j w_j^\intercal - \hat{W}_{j\to\mu}(t)\hat{W}_{j\to\mu}(t)^\intercal\end{cases}$$

In the limit $n\to\infty$, $X_{\mu i}$ being of order $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, the term I_j can be easily expanded and expressed with $\hat{W}$ and $\hat{C}$:

$$I_j = \int_{\mathbb{R}} dw_j m_{j\to\mu}(t,w_j)\exp\left(i\xi^\intercal X_{\mu j}w_j\right)) \simeq \exp\left(iX_{\mu j}\xi^\intercal\hat{W}_{j\to\mu}(t) - \frac{1}{2}X_{\mu j}^2\xi^\intercal\hat{C}_{j\to\mu}(t)\xi\right).$$

And finally using the inverse Fourier transform:

$$\begin{aligned}\tilde{m}_{\mu\to i}(t,w_i) &= \frac{1}{(2\pi)^K\mathcal{Z}_{\mu\to i}}\int_{\mathbb{R}^K} dz P_{\text{out}}(Y_\mu,z)\int_{\mathbb{R}^K} d\xi e^{-i\xi^\intercal z}e^{iX_{\mu i}\xi^\intercal w_i}\\ &\quad\times\prod_{j\neq i}^n \exp\left(iX_{\mu j}\xi^\intercal\hat{W}_{j\to\mu}(t) - \frac{1}{2}X_{\mu j}^2\xi^\intercal\hat{C}_{j\to\mu}(t)\xi\right)\\ &= \frac{1}{(2\pi)^K\mathcal{Z}_{\mu\to i}}\int_{\mathbb{R}^K} dz P_{\text{out}}(Y_\mu,z)\int_{\mathbb{R}^K} d\xi e^{-i\xi^\intercal z}e^{iX_{\mu i}\xi^\intercal w_i}e^{i\xi^\intercal\sum_{j\neq i}^n X_{\mu j}\hat{W}_{j\to\mu}(t)}e^{-\frac{1}{2}\xi^\intercal\sum_{j\neq i}^n X_{\mu j}^2\hat{C}_{j\to\mu}(t)\xi}\\ &= \frac{1}{(2\pi)^K\mathcal{Z}_{\mu\to i}}\int_{\mathbb{R}^K} dz P_{\text{out}}(Y_\mu,z)\sqrt{\frac{(2\pi)^K}{\det(V_{i\mu}^t)}}\underbrace{e^{-\frac{1}{2}\left(z-X_{\mu i}w_i-\omega_{i\mu}^t\right)^\intercal(V_{i\mu}^t)^{-1}\left(z-X_{\mu i}w_i-\omega_{i\mu}^t\right)}}_{:=H_{i\mu}}\end{aligned}$$

where we defined:

$$\begin{cases}\omega_{i\mu}^t \equiv \sum_{j\neq i}^n X_{\mu j}\hat{W}_{j\to\mu}(t)\\ V_{i\mu}^t \equiv \sum_{j\neq i}^n X_{\mu j}^2\hat{C}_{j\to\mu}(t)\end{cases}$$

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

$$\begin{aligned}H_{i\mu} \simeq e^{-\frac{1}{2}\left(z-\omega_{i\mu}^t\right)^\intercal(V_{i\mu}^t)^{-1}\left(z-\omega_{i\mu}^t\right)}\Bigg(1 + X_{\mu i}(w_i)^\intercal(V_{i\mu}^t)^{-1}(z-\omega_{i\mu}^t) - \frac{1}{2}X_{\mu i}^2(w_i)^\intercal(V_{i\mu}^t)^{-1}w_i\\ +\frac{1}{2}X_{\mu i}^2(w_i)^\intercal(V_{i\mu}^t)^{-1}(z-\omega_{i\mu}^t)(z-\omega_{i\mu}^t)^\intercal(V_{i\mu}^t)^{-1}w_i\Bigg).\end{aligned}$$

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

$$\begin{aligned}\tilde{m}_{\mu\to i}(t, w_i^2) \sim \frac{1}{\mathcal{Z}_{\mu\to i}} \Bigg\{ & 1 + X_{\mu i}(w_i)^\intercal g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) + \frac{1}{2} X_{\mu i}^2 (w_i)^\intercal g_{\text{out}} g_{\text{out}}^\intercal(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) w_i + \\ & \frac{1}{2} X_{\mu i}^2 (w_i)^\intercal \partial_\omega g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) w_i \Bigg\} \\ = \frac{1}{\mathcal{Z}_{\mu\to i}} & \left\{ 1 + (w_i)^\intercal B_{\mu\to i}^t + \frac{1}{2} (w_i)^\intercal B_{\mu\to i}^t (B_{\mu\to i}^t)^\intercal (w_i - \frac{1}{2} (w_i)^\intercal A_{\mu\to i}^t w_i \right\} \\ = & \sqrt{\frac{\det(A_{\mu\to i}^t)}{(2\pi)^K}} \exp\left(-\frac{1}{2} \left(w_i^\intercal - (A_{\mu\to i}^t)^{-1} B_{\mu\to i}^t \right)^\intercal A_{\mu\to i}^t \left(w_i^\intercal - (A_{\mu\to i}^t)^{-1} B_{\mu\to i}^t \right) \right)\end{aligned}$$

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

$$\begin{cases} B_{\mu\to i}^t \equiv X_{i\mu} g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) \\ A_{\mu\to i}^t \equiv -X_{i\mu}^2 \partial_\omega g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) \end{cases}$$

Using the set of BP equations (144), we can close the set of equations only over $\{m_{i\to\mu}\}_{i\mu}$:

$$m_{i\to\mu}(t+1, w_i) = \frac{1}{\mathcal{Z}_{i\to\mu}} P_0(w_i) \prod_{\nu\neq\mu}^m \sqrt{\frac{\det(A_{\nu\to i}^t)}{(2\pi)^K}} e^{-\frac{1}{2}\left(w_i - (A_{\nu\to i}^t)^{-1} B_{\nu\to i}^t\right)^\intercal A_{\nu\to i}^t \left(w_i - (A_{\nu\to i}^t)^{-1} B_{\nu\to i}^t\right)}.$$

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

$$\begin{cases} \Sigma_{\mu\to i}^t = \left(\sum_{\nu\neq\mu}^m A_{\nu\to i}^t \right)^{-1} \\ T_{\mu\to i}^t = \Sigma_{\mu\to i}^t \left(\sum_{\nu\neq\mu}^m B_{\nu\to i}^t \right) \end{cases} \qquad \begin{cases} \hat{W}_{i\to\mu}(t+1) = f_W(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) \\ \hat{C}_{i\to\mu}(t+1) = f_C(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) \end{cases}$$

Summary of the Relaxed BP set of equations

In the end, Relaxed BP equations are simply the following set of equations:

$$\begin{cases} \omega_{i\mu}^t = \sum_{j\neq i}^n X_{\mu j} \hat{W}_{j\to\mu}(t) \\ V_{i\mu}^t = \sum_{j\neq i}^n (X_{\mu j})^2 \hat{C}_{j\to\mu}(t) \\ B_{\mu\to i}^t = X_{\mu i} g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) \\ A_{\mu\to i}^t = -X_{\mu i}^2 \partial_\omega g_{\text{out}}(\omega_{i\mu}^t, Y_\mu, V_{i\mu}^t) \end{cases} \qquad \begin{cases} \Sigma_{\mu\to i}^t = \left(\sum_{\nu\neq\mu}^m A_{\nu\to i}^t \right)^{-1} \\ T_{\mu\to i}^t = \Sigma_{\mu\to i}^t \left(\sum_{\nu\neq\mu}^m B_{\nu\to i}^t \right) \\ \hat{W}_{i\to\mu}(t+1) = f_W(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) \\ \hat{C}_{i\to\mu}(t+1) = f_C(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) \end{cases} \tag{145}$$

G.2.2 Approximate Message Passing algorithm

On a tree, the missing message is negligible, which allows us to expand the previous Relaxed BP equations (145). We define the following estimates and parameters based on the complete set of messages:

$$\begin{cases} \omega_\mu^t = \sum_{j=1}^n X_{\mu j} \hat{W}_{j\to\mu}(t) \\ V_\mu^t = \sum_{j=1}^n X_{\mu j}^2 \hat{C}_{j\to\mu}(t) \end{cases} \qquad \begin{cases} \Sigma_i^t = \left(\sum_{\nu=1}^m A_{\nu\to i}^t \right)^{-1} \\ T_i^t = \Sigma_i^t \left(\sum_{\nu=1}^m B_{\nu\to i}^t \right) \end{cases} \tag{146}$$

- $\Sigma_{\mu\to i}^t$

$$\Sigma_{\mu\to i}^t = \left(\sum_{\nu\neq\mu}^m A_{\nu\to i}^t \right)^{-1} = \left(\sum_{\nu=1}^m A_{\nu\to i}^t - A_{\mu\to i}^t \right)^{-1} = \left(\sum_{\nu=1}^m A_{\nu\to i}^t \left(I_{K\times K} - \left(\sum_{\nu=1}^m A_{\nu\to i}^t \right)^{-1} A_{\mu\to i}^t \right) \right)^{-1}$$
$$= \left(I_{K\times K} - \left(\sum_{\nu=1}^m A_{\nu\to i}^t \right)^{-1} A_{\mu\to i}^t \right)^{-1} \left(\sum_{\nu=1}^m A_{\nu\to i}^t \right)^{-1} = \underbrace{\left(I_{K\times K} - \Sigma_i^t A_{\mu\to i}^t \right)^{-1}}_{\simeq I_{K\times K} + \Sigma_i^t A_{\mu\to i}^t + \mathcal{O}(n^{-1})} \Sigma_i^t \simeq \Sigma_i^t + \mathcal{O}(n^{-1})$$

- $T_{\mu\to i}^t$

$$T_{\mu\to i}^t = \Sigma_{\mu\to i}^t \left(\sum_{\nu\neq\mu}^m B_{\nu\to i}^t \right) = \left(\Sigma_i^t + \mathcal{O}\left(\frac{1}{n}\right) \right) \left(\sum_{\nu=1}^m B_{\nu\to i}^t - B_{\mu\to i}^t \right)$$
$$= T_i^t - \Sigma_i^t B_{\mu\to i}^t + \mathcal{O}\left(\frac{1}{n}\right)$$

- $\hat{W}_{i\to\mu}$

$$\hat{W}_{i\to\mu}(t+1) = f_W(\Sigma_{\mu\to i}^t, T_{\mu\to i}^t) = f_W\left(\Sigma_i^t, T_i^t - \Sigma_i^t B_{\mu\to i}^t\right) + \mathcal{O}\left(\frac{1}{n}\right)$$
$$\simeq f_W\left(\Sigma_i^t, T_i^t\right) - \left.\frac{df_W}{dT}\right|_{\left(\Sigma_i^t, T_i^t\right)} \Sigma_i^t B_{\mu\to i}^t$$
$$= \underbrace{f_W\left(\Sigma_i^t, T_i^t\right)}_{=\hat{W}_i(t+1)} - \left(\Sigma_i^t\right)^{-1} \underbrace{f_C\left(\Sigma_i^t, T_i^t\right) \Sigma_i^t}_{=\hat{C}_i(t+1)} \underbrace{B_{\mu\to i}^t}_{\simeq X_{\mu i} g_{\rm out}(\omega_\mu^t, Y_\mu, V_\mu^t)}$$
$$= \hat{W}_i(t+1) - X_{\mu i} \left(\Sigma_i^t\right)^{-1} \hat{C}_i(t+1) \Sigma_i^t g_{\rm out}(\omega_\mu^t, Y_\mu, V_\mu^t) + \mathcal{O}\left(\frac{1}{n}\right)$$

- $\hat{C}_{i\to\mu}$

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

$$\hat{C}_{i\to\mu}(t+1) = \mathbb{E}_{\tilde{P}_0}[\hat{W}_{i\to\mu} \hat{W}_{i\to\mu}^\intercal] - \mathbb{E}_{\tilde{P}_0}[\hat{W}_{i\to\mu}] \mathbb{E}_{\tilde{P}_0}[\hat{W}_{i\to\mu}]$$
$$= \mathbb{E}_{\tilde{P}_0}[\left(\hat{W}_i - X_{\mu i}\mathcal{E}\right) \left(\hat{W}_i - X_{\mu i}\mathcal{E}\right)^\intercal] - \mathbb{E}_{\tilde{P}_0}[\hat{W}_i - X_{\mu i}\mathcal{E}] \mathbb{E}_{\tilde{P}_0}[\hat{W}_i - X_{\mu i}\mathcal{E}]^\intercal$$
$$= \mathbb{E}_{\tilde{P}_0}[\hat{W}_i \hat{W}_i^\intercal] - \mathbb{E}_{\tilde{P}_0}[\hat{W}_i] \mathbb{E}_{\tilde{P}_0}[\hat{W}_i]^\intercal + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) = \hat{C}_i(t+1) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

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

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

- V_μ^t

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

- ω_μ^t

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

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

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

- T_i^t

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

The AMP algorithm follows naturally the rBP equations (145) using the expanded estimates of the mean and variance ω_μ,V_μ,T_i and Σ_i. The algorithm is written in pseudo language in Algorithm 1.

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

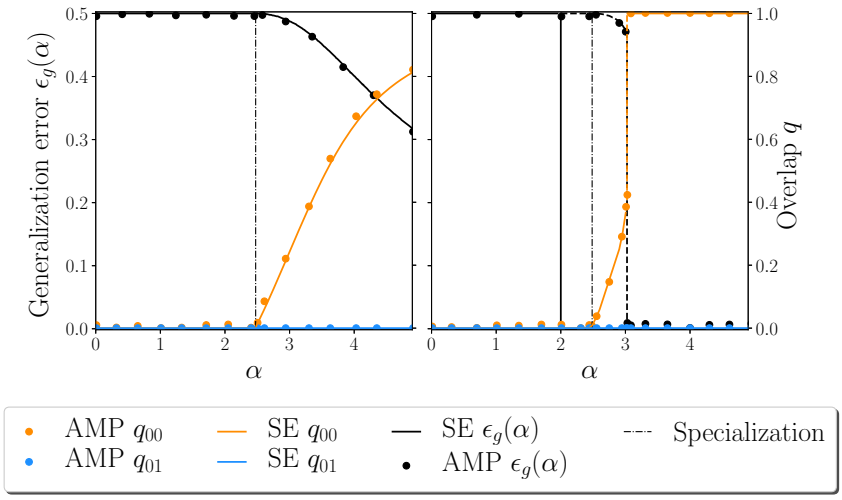

Figure 2: Similar plot as in Fig. 1 but for the parity machine with two hidden neurons. Value of the order parameter and the optimal generalization error for a parity machine with two hidden neurons with Gaussian weights (left) and binary/Rademacher weights (right). SE and AMP overlaps are respectively represented in full line and points.

the sign activation function has been replaced by $\varphi_1(z) = \mathbb{1}(z \neq 0) - \mathbb{1}(z = 0)$:

$$Y_\mu = \operatorname{sign}\Big[\prod_{l=1}^{K} \operatorname{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W_{il}^*\Big)\Big] = \varphi_1\Big[\sum_{l=1}^{K} \operatorname{sign}\Big(\sum_{i=1}^{n} X_{\mu i} W_{il}^*\Big)\Big]\,. \tag{147}$$

We have repeated our analysis for the $K = 2$ parity machine and the phase diagram is summarized in Fig. 2 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 $\alpha < \alpha_{\rm spec}$, the symmetry of the output imposes the non-specialized fixed point $q_{00} = q_{01} = 0$ to be the only solution, with $\alpha^G_{\rm spec}(K = 2) \simeq 2.48$ and $\alpha^B_{\rm spec}(K = 2) \simeq 2.49$. Above the specialization transition $\alpha_{\rm spec}$, 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 $\alpha^B_{\rm IT}(K = 2) \simeq 2.00$, 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 $\alpha^B_{\rm perf}(K = 2) \simeq 3.03$. 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.