Phase Transitions, Optimal Errors and Optimality of Message-Passing in Generalized Linear Models

Jean Barbier[†,♢], Florent Krzakala[⊗], Nicolas Macris[†], Léo Miolane[⋆,♢] and Lenka Zdeborová[*]

† Laboratoire de Théorie des Communications, Faculté Informatique et Communications,
Ecole Polytechnique Fédérale de Lausanne, 1015, Suisse.
⊗ Laboratoire de Physique Statistique, CNRS & Université Pierre et Marie Curie
& Ecole Normale Supérieure & PSL Université, Paris, France.
⋆ INRIA, 2 rue Simonne Iff, 75012, Paris, France.
∗ Institut de Physique Théorique, CNRS & CEA & Université Paris-Saclay, Saclay, France.
♢ Corresponding authors: jean.barbier@epfl.ch, leo.miolane@gmail.com

Abstract

We consider generalized linear models where an unknown n-dimensional signal vector is observed through the successive application of a random matrix and a non-linear (possibly probabilistic) componentwise function. We consider the models in the high-dimensional limit, where the observation consists of m points, and $m/n \to \alpha$ where α stays finite in the limit $m, n \to \infty$. This situation is ubiquitous in applications ranging from supervised machine learning to signal processing. A substantial amount of work suggests that both the inference and learning tasks in these problems have sharp intrinsic limitations when the available data become too scarce or too noisy. Here, we provide rigorous asymptotic predictions for these thresholds through the proof of a simple expression for the mutual information between the observations and the signal. Thanks to this expression we also obtain as a consequence the optimal value of the generalization error in many statistical learning models of interest, such as the teacher-student binary perceptron, and introduce several new models with remarquable properties. We compute these thresholds (or "phase transitions") using ideas from statistical physics that are turned into rigorous methods thanks to a new powerful smart-path interpolation technique called the stochastic interpolation method, which has recently been introduced by two of the authors. Moreover we show that a polynomial-time algorithm refered to as generalized approximate message-passing reaches the optimal generalization performance for a large set of parameters in these problems. Our results clarify the difficulties and challenges one has to face when solving complex high-dimensional statistical problems.

Contents

Authors contributions: J. Barbier, N. Macris and L. Miolane designed the proofs of the information theoretic results, namely of the replica formula and of the optimal errors. F. Krzakala and L. Zdeborová obtained the algorithmic and optimal phase transitions. All authors participated in the redaction of the paper.

I. Introduction

As datasets grow larger and more complex, modern statistical analysis and signal processing now requires solving very high-dimensional estimation problems with a very large number of parameters. This problematic arises in problems as diverse as deep learning [1] and regression problems [2] or compressed sensing in signal processing [3], [4]. Developing algorithms up to the task, and understanding their limitations, has become a major challenge in computer science, machine learning and statistics.

In many instances, it has been empirically observed that both the inference and learning tasks appear to have intrinsic limitations when the available data becomes too scarce or too noisy. In some cases these apparent thresholds are related to information theoretic phenomena: There is just not enough information in the dataset. This is the case, as famously discussed by Shannon in his seminal paper on communication theory [5], for the task of reconstructing a noisy signal when the noise is beyond the so-called Shannon capacity of the communication channel. In many situations, including communications, there also seem to exist jumps in the computational hardness, beyond which the most sophisticated known algorithms take exponential time to solve the task: In this case the problem has become too complicated to solve explicitly. A substantial amount of work suggests that one can understand and locate these fundamental barriers in many statistical models by thinking of them as *phase transitions* in the sense of physics. In fact, over the last three decades or so, a large body of interdisciplinary works in the statistical physics community has been applied to such problems [6]–[14] considering *random instances*, generated by given statistical models, and then locating these phase transitions. Such models are in fact being widely used in fields as diverse as (without any pretention at exhaustivity) statistical learning [15]–[19], compressed sensing and signal processing [14], [20]–[28], communication theory [29]–[33], community detection in networks [34]–[37], combinatorial optimization problems [13], [38], [39] or to model the behavior of neurons or neural nets [7], [8], [40], [41].

Many of these works, especially in the context of compressed sensing and machine learning, relied however on non-rigorous methods, using instead powerful heuristics like the replica and cavity approaches [6]. In the present contribution we leverage on these pioneering works and provide instead rigorous asymptotic predictions for several of these computational and information theoretic thresholds in the case of generalized linear estimation models [42]. This includes many popular statistical models of interests in many scientific fields such as random linear estimation in statistics, the teacher-student single perceptron problem, probit classification or quantized compressed sensing. For all these important estimation and learning problems our results completely vindicate the physics results —closing in some cases conjectures opened since almost three decades [43]–[45]— and considerably extend them in full generality. Additionally, we provide the value of the generalization error after an optimal learning, which gives a bound on how accurately *any algorithm* is able to predict outcome values for previously unseen data.

We also compare these optimal results with the algorithmic ones provided by message-passing algorihms [14], [21], [22] and observe that, while optimal performances are often thought to be intractable, they can actually be obtained using a polynomial-time scheme for a large set of parameters. There exists, however, an interesting region of parameters where all algorithms known to the authors fail to provide a satisfactory answer to the estimation and learning problems, while it is nevertheless information theoretically possible to do so. In this case there is a significant gap between what currently known polynomial algorithms can do and what should be expected from the information theoretic point of view.

Finally, our proof technique has an interest on its own. It exploits a powerful new technique called the stochastic interpolation method. It has been recently developed by two of the authors in [46] and is applicable to many other open problems in statistical estimation. Below we informally summarize our main contributions here:

- We consider generalized linear estimation models were, given an *unknown* signal vector $\mathbf{X}^*$, one is given the measurement vector $\mathbf{Y} = \varphi\big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]\big)$, with $\mathbf{\Phi}$ a known random matrix with i.i.d $\mathcal{N}(0,1)$ entries, and where φ acts componentwise.
- Our first main result is the rigorous determination of the expression of the (conditional) entropy $H(\mathbf{Y}|\mathbf{\Phi})$ of the observation variable, a quantity often called "the averaged free energy" in the statistical physics litterature, and this in the asymptotic limit where the number of variable is growing. As we shall see, many statistical quantities of interest, such as the mutual information between the observation and the unknown signal or the Bayes optimal generalization error, can be computed from this expression. We provide the proof of this expression —which is our main techniqueal contribution— in the last section. Only the simple linear case was known rigorously so far [47]–[49]. In fact, our results cover a large number of cases discussed in the litterature, often anticipated by statistical physics techniques, and allow to rigorously prove many predictions obtained by the heuristic replica method. The expression for the entropy was first famously conjectured in 1989 in [43] for the particular case of $\varphi(x) = \mathrm{sgn}(x)$, a work that is at the basis of many significant developments. The generic formula was also recently conjectured on the same heuristic basis [14]. Our proof yields a spectacular confirmation of the cavity and replica methods [6], [13].
- We compute the *Bayes-optimal generalization error* in the context of a *supervised learning* task of the rule used by the model (the so-called teacher-student problem). This estimator is optimal in the sense that it minimizes the label (the value of the ouput of φ) mean-square-error among all possible student estimators. Again, this formula can be anticipated with the heuristic cavity and replica methods, and was proposed, for the restricted case $\varphi(x) = \mathrm{sgn}(x)$ (the so-called perceptron problem), in pionnering works in the statistical physics community [7], [44], [50].
- We also compute (under some techniqueal hypotheses) the minimum mean-square-error (MMSE) for the reconstruction of the unknown signal in the generalized linear estimation model.
- While these results are information theoretic, we also consider the performance of a popular algorithm to solve random instances of generalized linear estimation problems, called generalized approximate message-passing (GAMP [14], [21], [22]). This algorithm, who also originated from statistical physics [40], [51]–[53], is expected to be particularly powerful on these random instances, as proven for instance in compressed sensing [21], [54], [55]. Indeed we show that, for a large set of problems and a large region of parameters, GAMP yields optimal generalization error. It exists, however, a region where GAMP does not reach the optimal results. In this case, we conjecture that the algorithmic problem is computationally hard.
- Finally, we study in depth the situation for many given choices of the function φ, and identify sharp phase transitions and novel phase transitions. By locating these transitions we clarify the difficulties and challenges in solving complex non-linear high-dimensional statistical problems in many concrete situations, and characterize these problems in terms of optimal information theoretical reconstruction.

II. Setting and main results

A. Generalized linear estimation: Problem statement

We now a generic description of the observation model. Note that we describe here an *estimation* (or inference) problem. In the title of the paper, we refer to generalized linear *models*. This is because the setting that we describe now is very generic and will allow us to also consider supervised *learning* problems (see Sec. III). Precise hypotheses are given below in Sec. II-C.

Let $n, m \in \mathbb{N}^*$. Let P_0 be a probability distribution over $\mathbb{R}$ and let $X_1^*, \dots, X_n^* \overset{\text{iid}}{\sim} P_0$ be the components of a signal vector $\mathbf{X}^*$ (this is also denoted $\mathbf{X}^* \overset{\text{iid}}{\sim} P_0$). We fix a function $\varphi : \mathbb{R} \times \mathbb{R}^{k_A} \to \mathbb{R}$ and consider $(A_\mu)_{\mu=1}^m \overset{\text{iid}}{\sim} P_A$, where P_A is a probability distribution over $\mathbb{R}^{k_A}$ ($k_A \in \mathbb{N}$). We acquire m measurements through

$$Y_\mu = \varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu, A_\mu\Big) + \sqrt{\Delta} Z_\mu\,, \qquad 1 \le \mu \le m\,, \tag{1}$$

where $Z_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ is an additive Gaussian noise, $\Delta > 0$, and $\mathbf{\Phi}$ is a $m \times n$ measurement matrix with i.i.d $\Phi_{\mu i} \sim \mathcal{N}(0,1)$ entries. The estimation problem is to recover $\mathbf{X}^*$ from the knowledge of $\mathbf{Y} = (Y_\mu)_{\mu=1}^m$, φ, $\mathbf{\Phi}$, Δ, P_0 and P_A (the realization of the random stream $\mathbf{A}$ itself, if present in the model, is unknown). We use the notation $[\mathbf{\Phi X}^*]_\mu = \sum_{i=1}^n \Phi_{\mu i} X_i^*$. When $\varphi(x, A) = x$ we have a random linear estimation problem, whereas if, say, $\varphi(x, A) = \mathrm{sgn}(x)$ we have a noisy single layer perceptron. Sec. III discusses various examples related to non-linear estimation and supervised learning.

It also fruitful to think of the measurements as the outputs of a "channel",

$$Y_\mu \sim P_{\rm out}\Big(\,\cdot\,\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu\Big) \tag{2}$$

where the transition density (with respect to Lebesgue's measure) is

$$P_{\rm out}\Big(y_\mu\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu\Big) = \frac{1}{\sqrt{2\pi\Delta}}\int dP_A(a_\mu) e^{-\frac{1}{2\Delta}\left(y_\mu - \varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu, a_\mu)\right)^2}\,. \tag{3}$$

Our hypotheses ensure that this is a well defined density. In fact (3) is sometimes called a "random function representation" of a transition kernel. Our analysis uses both representations (1) and (2).

Throughout this paper we often adopt the language of statistical mechanics. In particular the random variables $\mathbf{Y}$ (and also $\mathbf{\Phi}$, $\mathbf{X}^*$, $\mathbf{A}$, $\mathbf{Z}$) are called *quenched* variables because once the measurements are acquired they have a "fixed realization." An expectation taken with respect to *all* quenched r.v appearing in an expression will simply be denoted by $\mathbb{E}$ *without* subscript. Subscripts are only used when the expectation carries over a subset of r.v appearing in an expression or when some confusion could arise.

A fundamental role is played by the joint posterior distribution of (the signal) $\mathbf{x}$ and of (the random stream) $\mathbf{a}$ given the quenched measurements $\mathbf{Y}$ (recall that both $\mathbf{X}^*$ and $\mathbf{A}$ are unknown). The prior over the signal is denoted $dP_0(\mathbf{x}) = \prod_{i=1}^n dP_0(x_i)$, and similarly $dP_A(\mathbf{a}) = \prod_{\mu=1}^m dP_A(a_\mu)$. According to the Bayes formula this joint posterior is given by

$$dP(\mathbf{x} = \mathbf{X}^*, \mathbf{a} = \mathbf{A}|\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})} dP_0(\mathbf{x}) dP_A(\mathbf{a}) \prod_{\mu=1}^m \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(Y_\mu - \varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu, a_\mu)\right)^2}\,, \tag{4}$$

where the *partition function* (the normalization factor) is defined as

$$\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi}) := \int dP_0(\mathbf{x}) dP_A(\mathbf{a}) \prod_{\mu=1}^m \frac{1}{\sqrt{2\pi\Delta}} e^{-\frac{1}{2\Delta}\left(Y_\mu - \varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu, a_\mu)\right)^2}\,. \tag{5}$$

Marginalizing (4) w.r.t $\mathbf{a}$ leads the posterior of $\mathbf{x}$, namely

$$dP(\mathbf{x} = \mathbf{X}^*|\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})} dP_0(\mathbf{x}) e^{-\mathcal{H}(\mathbf{x}; \mathbf{Y}, \mathbf{\Phi})}\,, \tag{6}$$

$$\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi}) = \int dP_0(\mathbf{x}) e^{-\mathcal{H}(\mathbf{x}; \mathbf{Y}, \mathbf{\Phi})}\,. \tag{7}$$

where the *Hamiltonian* is defined as

$$\mathcal{H}(\mathbf{x}; \mathbf{Y}, \mathbf{\Phi}) := -\sum_{\mu=1}^m \ln P_{\rm out}\Big(Y_\mu\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu\Big)\,. \tag{8}$$

From the point of view of statistical mechanics (6) is a Gibbs distribution and the integration over $dP_0(\mathbf{x})$ in the partition function is best thought as a "sum over annealed or fluctuating degrees of freedom" (note that in the representation (5), $(a_\mu)_{\mu=1}^m$ also play the role of annealed variables). Let us introduce a standard statistical mechanics notation for the expectation w.r.t the join posterior (4), the so called *Gibbs bracket* $\langle - \rangle$ defined as

$$\langle g(\mathbf{x}, \mathbf{a}) \rangle := \int dP(\mathbf{x} = \mathbf{X}^*, \mathbf{a} = \mathbf{A}|\mathbf{Y}, \mathbf{\Phi}) g(\mathbf{x}, \mathbf{a}) \tag{9}$$

for any function g such that this expectation exists.

The main quantity of interest here is the associated *free entropy* (or minus the *free energy*)

$$f_n := \frac{1}{n}\mathbb{E} \ln \mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})\,. \tag{10}$$

It is perhaps useful to stress that $\mathcal{Z}(\mathbf{Y}, \mathbf{\Phi})$ is nothing else than the density of $\mathbf{Y}$ conditioned on $\mathbf{\Phi}$ so we have the explicit representation (used later on)

$$f_n = \frac{1}{n}\mathbb{E}_{\mathbf{\Phi}} \int d\mathbf{Y} \mathcal{Z}(\mathbf{Y}, \mathbf{\Phi}) \ln \mathcal{Z}(\mathbf{Y}, \mathbf{\Phi}) = \frac{1}{n}\mathbb{E}_{\mathbf{\Phi}} \int d\mathbf{Y} dP_0(\mathbf{X}^*) e^{-\mathcal{H}(\mathbf{X}^*; \mathbf{Y}, \mathbf{\Phi})} \ln \int dP_0(\mathbf{x})\, e^{-\mathcal{H}(\mathbf{x}; \mathbf{Y}, \mathbf{\Phi})}\,, \tag{11}$$

where $d\mathbf{Y} = \prod_{\mu=1}^m dY_\mu$. Thus f_n is minus the conditional entropy $-H(\mathbf{Y}|\mathbf{\Phi})/n$ of the measurements. One of the main contributions of this paper is the derivation, thanks to the stochastic interpolation method, of the thermodynamic limit $\lim_{n\to\infty} f_n$ in the "high-dimensional" regime, namely when $n, m \to \infty$ while $m/n \to \alpha > 0$ (α is sometimes refered to as the "measurement rate" in compressed sensing terminology).

B. Two scalar inference channels

An important role in our proof of the asymptotic expression of the free entropy is played by simple *scalar* inference channels. As we will see, the free entropy is expressed in terms of the free entropy of these channels. This "decoupling property" stands at the root of the mean-field approach in statistical physics, used through in replica method to perform a formal calculation of the free entropy of the model [6], [13]. Let us now introduce these two scalar denoising models.

The first one is an additive Gaussian channel. Let $r \geq 0$, which play the role of a signal-to-noise ratio (snr). Suppose that $X_0 \sim P_0$ and that we observe

$$Y_0 = \sqrt{r}\, X_0 + Z_0 \,, \tag{12}$$

where $Z_0 \sim \mathcal{N}(0,1)$ independently of X_0. Consider the inference problem consisting of retrieving X_0 from the observations Y_0. The associated posterior distribution is

$$dP(x = X_0|Y_0) = \frac{dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2}}{\int dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2}} \,. \tag{13}$$

In this expression all the x-independent terms have been simplified between the numerator and the normalization. The free entropy associated with this channel is just the expectation of the logarithm of the normalization factor

$$\psi_{P_0}(r) := \mathbb{E} \ln \int dP_0(x) e^{\sqrt{r}\, Y_0 x - r x^2/2} \,. \tag{14}$$

The second scalar channel that appears naturally in the problem is linked to the kernel $P_{\rm out}$ through the following inference model. Suppose that $V, W^* \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ where V is *known* while the inference problem is to recover the unknown W^* from the following observation

$$\widetilde{Y}_0 \sim P_{\rm out}\big(\,\cdot\,|\sqrt{q}\, V + \sqrt{\rho - q}\, W^*\big) \,, \tag{15}$$

where $\rho > 0$ and $q \in [0, \rho]$. The free entropy for this model, again related to the normalization factor of the posterior $dP(w = W^*|\widetilde{Y}_0, V)$, is

$$\Psi_{P_{\rm out}}(q;\rho) = \Psi_{P_{\rm out}}(q) := \mathbb{E} \ln \int dw \frac{e^{-\frac{w^2}{2}}}{\sqrt{2\pi}} P_{\rm out}\big(\widetilde{Y}_0|\sqrt{q}\, V + \sqrt{\rho - q}\, w\big) \,. \tag{16}$$

We prove in Appendix B that this function is twice differentiable and convex with respect to (w.r.t) its first argument.

C. Replica-symmetric formula, mutual information and optimal output error

Let us now introduce our first main result, namely a complete proof of the single-letter *replica-symmetric formula* for the asymptotic free entropy of model (1), (2). The proof is performed under the following rather general hypotheses.

(h1) The prior distribution P_0 admits a finite second moment.
(h2) For some $\gamma > 0$ the moment of order $2 + \gamma$ of $|\varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1, A_1)|$ is bounded uniformly in n.

For concreteness the reader can keep in mind the class of polynomially bounded measurement models such that $\varphi(z,a) \leq c_1 + c_2|z|^p$ for some constants $c_1 > 0$, $c_2 > 0$, $p \geq 1$. In Appendix C we verify that (h2) is satisfied for such measurements as long as P_0 has finite $p(2+\gamma)$-th moments. Notice that no continuity or differentiability assumption on φ is required.

Let us define the *replica-symmetric potential* (or just potential). Call $\rho := \mathbb{E}[(X^*)^2]$ where $X^* \sim P_0$. Then the potential is

$$f_{\rm RS}(q,r;\rho) = f_{\rm RS}(q,r) := \psi_{P_0}(r) + \alpha \Psi_{P_{\rm out}}(q;\rho) - \frac{rq}{2} \,. \tag{17}$$

From now on denote $\psi'_{P_0}(r)$ and $\Psi'_{P_{\rm out}}(q) = \Psi'_{P_{\rm out}}(q;\rho)$ the derivatives of $\psi_{P_0}(r)$ and $\Psi_{P_{\rm out}}(q;\rho)$ w.r.t their first argument. The main theorem of this paper is

***Theorem** 2.1 (Replica-symmetric formula):* For the generalized estimation model (1), (2) and under the hypotheses (h1), (h2) the thermodynamic limit of the free entropy (10) verifies

$$f_\infty := \lim_{n\to\infty} f_n = \sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}(q,r) = \sup_{(q,r)\in\Gamma} f_{\rm RS}(q,r) \,, \tag{18}$$

where the elements of Γ are called "fixed points of the state evolution", and are defined by:

$$\Gamma := \left\{ (q,r) \in [0,\rho] \times \mathbb{R}_+ \,\middle|\, \begin{array}{rcl} q & = & 2\psi'_{P_0}(r) \\ r & = & 2\alpha \Psi'_{P_{\rm out}}(q;\rho) \end{array} \right\} . \tag{19}$$

Moreover, the "$\sup\inf$" and the supremum over Γ in (18) are achieved over the same couples.

The theorem will first be proved under the simpler assumptions of P_0 with bounded support and φ bounded, twice differentiable with respect to its first argument, with bounded first and second derivative. In Appendix F we give approximation arguments to cover models satisfying (h1) and (h2).

An imediate corollary of Theorem 2.1 is the limiting expression of the mutual information between the observations and the hidden variables.

***Corollary** 2.2 (Single-letter formula for the mutual information):* The thermodynamic limit of the mutual information for model(1), (2) between the observations and the hidden variables verifies

$$i_n := \frac{1}{n} I(\mathbf{X}^*, \mathbf{A}; \mathbf{Y}|\mathbf{\Phi}) \xrightarrow[n\to\infty]{} i_\infty := -f_\infty - \frac{\alpha}{2}(1+\ln(2\pi\Delta))\,. \tag{20}$$

Proof: This follows from a simple calculation:

$$\begin{aligned}\frac{1}{n} I(\mathbf{X}^*, \mathbf{A}; \mathbf{Y}|\mathbf{\Phi}) &= \mathbb{E}\ln\frac{P(\mathbf{Y},\mathbf{X}^*,\mathbf{A}|\mathbf{\Phi})}{P(\mathbf{Y}|\mathbf{\Phi})P(\mathbf{X}^*,\mathbf{A}|\mathbf{\Phi})} = -\frac{1}{n}\mathbb{E}\ln P(\mathbf{Y}|\mathbf{\Phi}) + \frac{1}{n}\mathbb{E}\ln P(\mathbf{Y}|\mathbf{X}^*,\mathbf{A},\mathbf{\Phi}) & (21)\\ &= -f_n - \frac{1}{2n\Delta}\mathbb{E}\sum_{\mu=1}^m (Y_\mu - \varphi([\mathbf{\Phi}\mathbf{X}^*]_\mu, A_\mu))^2 - \frac{m}{2n}\ln(2\pi\Delta) & (22)\\ &= -f_n - \frac{m}{2n} - \frac{m}{2n}\ln(2\pi\Delta)\,. & (23)\end{aligned}$$

■

Another corollary is the following expression for the "measurement (or output) minimum-mean-square error". Let us define a Gibbs bracket for the scalar channel at fixed V:

$$\langle g(w,a)\rangle_{\rm sc} = \int d\widetilde{Y}_0 \mathcal{D}w dP_A(a) \frac{e^{-\frac{1}{2\Delta}\big(\widetilde{Y}_0-\varphi(\sqrt{q}\,V+\sqrt{\rho-q}\,w,a)\big)^2}}{\int \mathcal{D}w dP_A(a) e^{-\frac{1}{2\Delta}\big(\widetilde{Y}_0-\varphi(\sqrt{q}\,V+\sqrt{\rho-q}\,w,a)\big)^2}} g(w,a)\,, \tag{24}$$

where $\mathcal{D}w = dw(2\pi)^{-1/2}e^{-w^2/2}$ is a standard Gaussian measure.

***Corollary** 2.3 (Single-letter formula for the output minimum-mean-square error):* For almost every $\Delta > 0$, for any optimal couple (q^*, r^*) of (18) we have

$$\begin{aligned}&\frac{1}{2}\lim_{n\to\infty}\frac{1}{m}\mathbb{E}\Big\langle\Big\|\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}\mathbf{X}^*,\mathbf{A}\Big)-\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}\mathbf{x},\mathbf{a}\Big)\Big\|^2\Big\rangle\\ &=\lim_{n\to\infty}\frac{1}{m}\mathbb{E}\Big[\Big\|\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}\mathbf{X}^*,\mathbf{A}\Big)-\Big\langle\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}\mathbf{x},\mathbf{a}\Big)\Big\rangle\Big\|^2\Big]\\ &=\frac{1}{2}\mathbb{E}\big\langle\big(\varphi(\sqrt{q^*}\,V+\sqrt{\rho-q^*}\,W^*,A)-\varphi(\sqrt{q^*}\,V+\sqrt{\rho-q^*}\,w,a)\big)^2\big\rangle_{\rm sc}\\ &=\mathbb{E}\big[\varphi(\sqrt{\rho}\,V,A)^2\big]-\mathbb{E}\big[\langle\varphi(\sqrt{q^*}\,V+\sqrt{\rho-q^*}\,w,a)\rangle_{\rm sc}^2\big]\,,\end{aligned} \tag{25}$$

where the Gibbs brackets are defined by (9) and (24) and $V, W^* \overset{\rm iid}{\sim} \mathcal{N}(0,1)$, $A\sim P_A$.

Proof: The first equality is a direct consequence of the *Nishimori identity* (a fundamental identity that follows directly from Bayes formula and that will play a major role in our proofs, see Appendix A). This precise sub-identity, as well as its proof, can be found in Appendix B of [48] for the linear case. This sub-identity also allows to prove the equality between the third and last terms of (25).

The proof of the equality between the first and third terms of (25) works as follows. One can verify easily that i_n is a concave differentiable function of Δ^{-1} (see [48] for such a proof). Thus its limit i_∞ is also a concave function of Δ^{-1}. Therefore, a standard analysis lemma gives that the derivative of i_n w.r.t Δ^{-1} converges to the derivative of i_∞ at every point at which i_∞ is differentiable (i.e. almost every points, by concavity). Let us now compute these two derivatives. First, using Gaussian integration by parts (using the elementary formula $\mathbb{E}_Z[Zg(Z)]=\mathbb{E}_Z[\partial_x g(x)|_{x=Z}]$ for $Z\sim\mathcal{N}(0,1)$) and the Nishimori identity, one can verify the following (generalized) I-MMSE relation (see Appendix A in [48] for a similar proof)

$$\frac{\partial i_n}{\partial\Delta^{-1}} = \frac{1}{4n}\sum_{\mu=1}^m\mathbb{E}\Big\langle\Big(\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu,A_\mu\Big)-\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{x}]_\mu,a_\mu\Big)\Big)^2\Big\rangle\,. \tag{26}$$

Second, define

$$h(q,\Delta) = \frac{\alpha}{2}(1+\ln(2\pi\Delta)) + \alpha\Psi_{P_{\rm out}}(q) + \inf_{r\geq 0}\Big\{\psi_{P_0}(r) - \frac{qr}{2}\Big\}\,. \tag{27}$$

Then $i_\infty = -\sup_{q\in[0,\rho]} h(q,\Delta)$. Compute (again, using Gaussian integration by parts and the Nishimori identity) for $\Delta > 0$ and $q \in [0,\rho]$:

$$\frac{\partial h}{\partial \Delta^{-1}}(q,\Delta) = \alpha \frac{\partial \Psi_{P_{\rm out}}}{\partial \Delta^{-1}}(q,\Delta) - \alpha\Delta = -\frac{\alpha}{4}\mathbb{E}\big\langle \big(\varphi(\sqrt{q}\,V + \sqrt{\rho-q}\,W^*, A) - \varphi(\sqrt{q}\,V + \sqrt{\rho-q}\,w, a)\big)^2\big\rangle_{\rm sc}\,. \tag{28}$$

Theorem 1 from [56] gives that at every Δ^{-1} at which i_∞ is differentiable

$$\frac{\partial i_\infty}{\partial \Delta^{-1}} = -\frac{\partial}{\partial \Delta^{-1}} \sup_{q\in[0,\rho]} h(q,\Delta) = \frac{\alpha}{4}\mathbb{E}\big\langle \big(\varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,W^*, A) - \varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)\big)^2\big\rangle_{\rm sc}\,. \tag{29}$$

As explained above, $\partial i_n / \partial \Delta^{-1}$ converges to $\partial i_\infty / \partial \Delta^{-1}$ at every Δ^{-1} at which i_∞ is differentiable, which concludes the proof. ■

As it will become clear in Sec. III this output error is related to the optimal learning (or training) error and optimal generalization error in a supervised learning setting, and thus plays a fundamental role.

D. Optimality of the generalized approximate message-passing algorithm

Another main result of the paper is a simple expression for the Bayes-optimal generalization errror. We refer to Sec. III for a precise definition of this error and the associated formula.

While the main results presented until now are information theoretic, our last one concerns the performance of a popular algorithm to solve random instances of generalized linear problems, called generalized ppproximate message-passing (GAMP). We shall not re-derive its properties here, and instead refer to the original papers for details. This approach has a long history, especially in statistical physics [40], [51]–[53], error correcting codes [57], and graphical models [58]. For a modern derivation in the context of linear models, see [21], [54], [55]. The case of generalized linear models was discussed by Rangan in [22], and has been used for classifcation purpose in [59].

Define the so-called threshold function $\eta(\Sigma, R)$ as the expectation of the variable x sampled from the following distribution $C\,P_0(x)\exp\{-(R-x)^2/(2\Sigma)\}$ ($C = C(\Sigma,R)$ is the normalization). Moreover we need to define the so-called output function $g_{\rm out}(\omega, Y, V) = \partial_\omega \ln \int dz P_{\rm out}(Y|z)\exp\{-(z-\omega)^2/(2V)\}/\sqrt{2\pi V}$. This function acts componentwise when applied to vectors. Given initial estimates $\mathbf{a}^0$, $\mathbf{v}^0$ for the means and variances of the elements of the signal vector $\mathbf{X}^*$, GAMP takes as input the observation vector $\mathbf{Y}$ and then iterates the following equations with initialization $g^0_\mu = 0$ for all $\mu = 1,\dots,m$ (we denote by $\overline{\mathbf{u}}$ the average over all the components of the vector $\mathbf{u}$ and $\mathbf{\Phi}^\intercal$ is the transpose of the matrix $\mathbf{\Phi}$): From $t = 1$ until convergence,

$$\begin{cases} V^t &=& \overline{\mathbf{v}^t} \\ \boldsymbol{\omega}^t &=& \mathbf{\Phi}\mathbf{a}^{t-1}/\sqrt{n} - V^t\mathbf{g}^{t-1} \\ g^t_\mu &=& g_{\rm out}(\omega^t_\mu, Y_\mu, V^t) \quad \forall\ \mu = 1,\dots m \\ \Sigma^t &=& \big(\alpha\, \overline{g^2_{\rm out}(\boldsymbol{\omega}^t, \mathbf{Y}, V^t)}\big)^{-1} \\ \mathbf{R}^t &=& \mathbf{a}^{t-1} + \mathbf{\Phi}^\intercal\mathbf{g}^t/(\Sigma^t\sqrt{n}) \\ a^t_i &=& \eta(\Sigma^t, R^t_i) \quad \forall\ i = 1,\dots n \\ v^t_i &=& \Sigma^t\, \partial_R\eta(\Sigma^t, R)|_{R=R^t_i} \quad \forall\ i = 1,\dots n \end{cases} \tag{30}$$

One of the strongest asset of GAMP is that its performance can be tracked rigorously in the limit $n, m \to \infty$ via a procedure known as state evolution (SE), see [60], [61] for the linear case, and [22], [62] for the generalized one. In our notations, state evolution tracks the asymptotic value of the overlap between the true hidden value $\mathbf{X}^*$ and its estimate by GAMP $\mathbf{a^t}$ defined as $m^t = \lim_{n\to\infty} \mathbf{X}^*{\cdot}\mathbf{a}^t/n$ (that is related to the asymptotic mean-square error E^t between $\mathbf{X}^*$ and its estimate $\mathbf{a}^t$ by $E^t = \rho - m^t$, where recall that $\rho = \mathbb{E}[(X^*)^2]$ with $X^* \sim P_0$) via:

$$m^{t+1} = \psi'_{P_0}(\widehat{m}^t)/2\,, \tag{31}$$

$$\widehat{m}^t = \alpha\Psi'_{P_{\rm out}}(m^t;\rho)/2\,. \tag{32}$$

From Theorem 2.1 we realize that the fixed points of these equations correspond to the critical point of the asymptotic conditional entropy in (18) where (q,r) are the equivalent of $(m,\widehat{m})$. In fact, in the replica heuristic, the extremizer q^* is conjectured to give the optimal value of the overlap, a fact that was proven rigorously for the linear channel [47]. If the minimizer of (18) is unique, or is attractive whithin the GAMP dynamics, then m^t converges to q^* at long time.

Perhaps more surprisingly, one can use GAMP in the teacher-student scenario (that we precisely describe in Sec. III-A) in order to provide an estimation of a new output $C_{\rm new} = \varphi(\mathbf{\Phi}_{\rm new}\cdot\mathbf{X}^*/\sqrt{n}, A_{\rm new})$, where $\mathbf{\Phi}_{\rm new}$ is a new row of the matrix and $A_{\rm new} \sim P_A$ a new random number. Let $\mathbf{x}$ be drawn according to the true posterior (6). As $\mathbf{a}^t$ is the GAMP estimate of the expectation of $\mathbf{x}$, with estimated variance V^t, the natural heuristic is to consider for the posterior probability distribution of the random variable $w := \mathbf{\Phi}_{\rm new} \cdot \mathbf{x}/\sqrt{n}$ a Gaussian with mean $\mathbf{\Phi}_{\rm new} \cdot \mathbf{a}^t/\sqrt{n}$ and variance V^t. This allows to estimate the posterior mean of the output, which leads the GAMP prediction (recall the channel $P_{\rm out}$ definition (3)):

$$\widehat{C}^{\rm GAMP,\mathit{t}} = \int dy\, dw\, y\, P_{\rm out}(y|w) \frac{1}{\sqrt{2\pi V^t}} e^{-\frac{1}{2V^t}\left(\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\rm new}\cdot \mathbf{a}^t - w\right)^2} . \tag{33}$$

A straightforward application of the state evolution analysis in [62] then indicates that this rigorously leads to a GAMP generalization error given by

$$\mathcal{E}_{\rm gen}^{\rm GAMP,\mathit{t}} := \lim_{n\to\infty} \mathbb{E}\big[\big(C_{\rm new} - \widehat{C}^{\rm GAMP,\mathit{t}}\big)^2\big] = \mathbb{E}_{V,a}\big[\varphi(\sqrt{\rho}\, V, a)^2\big] - \mathbb{E}_V\big[\mathbb{E}_{w,a}\big[\varphi(\sqrt{m^t}\, V + \sqrt{\rho - m^t}\, w, a)\big]^2\big] , \tag{34}$$

where $V, w \overset{\rm iid}{\sim} \mathcal{N}(0,1)$ and $a \sim P_A$ (do not get confused between V which is a dummy r.v and V^t the variance of the GAMP estimate, and between the dummy r.v a and the GAMP estimate $\mathbf{a}^t$).

We will see in Sec. III-B that this formula matches the one for the Bayes-optimal generalization error, see (38), up to the fact that instead of q^* (a maximizer obtained from the replica formula (18)) appearing in the optimal error formula, here it appears m^t. Thus clearly, when m^t converges to q^* (we shall see that this is the case in many situations in the examples of Sec. IV) this yields a very interesting and non trivial result: *GAMP achieves the Bayes-optimal generalization error* in a plethora of models (a task again often believed to be intractable) and this for large sets of parameters.

III. Optimal generalization error in supervised learning

In this section we show how Theorem 2.1 allows to compute the optimal generalization and learning (or training) errors. We will then apply our findings to specific examples in Sec. IV.

The goal here is to develop a conceptual framework for deriving both error expressions but, as we will see, the final expressions correspond simply to the single letter formula of the output MMSE (the last expression in Corollary 2.3) but considered in two different regimes: For obtaining the learning error we need to consider a finite noise Δ, which corresponds to the noise in the data used during the learning stage, and q^* a maximizer of the replica formula (18) (evaluated at this finite noise Δ). Instead, considering the large $\Delta \to \infty$ expansion of the output MMSE (25) (this is done in Appendix J-B) and plugging inside it the *same* q^* as for the learning error we obtain the generalization error. The q^* used for computing the generalization error is the same as for the learning error because, informally speaking, this takes into account the information gained from the learning stage that allows to learn the model and thus to estimate new outputs.

Let us start by explaining how the *inference* problem (1) can be re-interpreted as a supervised *learning* problem.

A. Teacher-student scenario

Consider the following *teacher-student scenario* (also called planted model). We voluntarily employ terms coming from machine learning, instead of the signal processing terminology used in the previous sections.

First the teacher randomly generates a *classifier* $\mathbf{X}^* \in \mathbb{R}^n$ (the signal in the estimation problem) with $\mathbf{X}^* \overset{\rm iid}{\sim} P_0$, an ensemble of $m = \alpha n$ *patterns* (row-vectors) $\mathbf{\Phi}_\mu \in \mathbb{R}^n$ for $\mu = 1, \dots, m$ and such that $\mathbf{\Phi}_\mu \overset{\rm iid}{\sim} \mathcal{N}(0,1)$, and a stream $\mathbf{A} = (A_\mu)_{\mu=1}^m$ with $\mathbf{A} \overset{\rm iid}{\sim} P_A$. The teacher chooses a model $\varphi(x, A)$. For deterministic models A is simply absent. Finally the teacher associates to each pattern a *label* $C_\mu \in \mathbb{R}$ selected by the classifier, namely $C_\mu = \varphi(\mathbf{\Phi}_\mu \cdot \mathbf{X}^*/\sqrt{n}, A_\mu)$ for $\mu = 1, \dots, m$. Stacking the rows $\{\mathbf{\Phi}_\mu\}_{\mu=1}^m$ in a $m \times n$ matrix $\mathbf{\Phi}$ and denoting the vector of labels $\mathbf{C} = (C_\mu)_{\mu=1}^m$, then the labels, patterns, source of randomness and classifier verify

$$\mathbf{C} = \varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}\mathbf{X}^*, \mathbf{A}\Big) , \tag{35}$$

where it is understood that φ acts in componentwise fashion on vectors. The student is given the distributions P_0, P_A and the function φ and his task is to learn the classifier $\mathbf{X}^*$ from a subset of the pattern-label pairs.

More precisely we consider the following scenario. The set of rows of $\mathbf{\Phi}$ and labels are divided into two sets by the teacher: The *training set* $\mathcal{S}^{\rm tr}$ of size βm, $\beta \in [0,1]$, that will be used by the student in order to learn the classifier, and the *test set* $\mathcal{S}^{\rm te}$ of size $(1-\beta)m$ that will be used by the teacher in order to evaluate the performance of the student. For the training set both the patterns *and* associated noisy labels are given to the student, namely $\mathcal{S}^{\rm tr} = \{(Y_\mu = C_\mu + Z_\mu\sqrt{\Delta^{\rm tr}}; \mathbf{\Phi}_\mu)\}_{\mu=1}^{\beta m}$ where $Z_\mu \sim \mathcal{N}(0,1)$. Here $\Delta^{\rm tr}$ is strictly positive (but typically small) and known by the student who can thus optimally tune its "confidence" in the training labels during the training/learning stage. For the test set, the previously unseen patterns to classify are given to the student but the labels are not, namely $\mathcal{S}^{\rm te} = \{(Y_\mu = C_\mu + Z_\mu\sqrt{\Delta^{\rm te}}; \mathbf{\Phi}_\mu)\}_{\mu=\beta m+1}^{m}$ where $\Delta^{\rm te} \to \infty$ (for the derivation of the errors, we first consider a finite $\Delta^{\rm te}$ and we will then let it diverge to obtain the final expression of the generalization error).

Define $\Delta_\mu = \Delta^{\rm tr}$ if $\mu \le \beta m$, $\Delta_\mu = \Delta^{\rm te}$ else. Then the inference of the classifier $\mathbf{X}^*$ from the noisy labels $Y_\mu = C_\mu + Z_\mu\sqrt{\Delta_\mu}$ is a slight extension of model (1): This inference problem is nothing else than a *supervised learning* of the classifier by the student.

B. Optimal generalization error

An important quantity is the *generalization error* which measures the performance of the student. Define $\mathbf{C}^{\rm te} = (C_u)_{\mu=\beta m+1}^m$ and $\mathbf{\Phi}^{\rm te} = \{\mathbf{\Phi}_\mu\}_{\mu=\beta m+1}^m$ as the $m(1-\beta)$-dimensional vector of labels and the $m(1-\beta)\times n$ matrix, respectively, both restricted to the *test set* and similarly for $\mathbf{A}^{\rm te} = (A_\mu)_{\mu=\beta m+1}^m$. Then we define the generalization error at finite $\Delta^{\rm te}$ as

$$\mathcal{E}_{\rm gen} := \frac{1}{(1-\beta)m}\mathbb{E}\Big[\Big\|\mathbf{C}^{\rm te} - \Big\langle\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm te}\mathbf{x}, \mathbf{a}^{\rm te}\Big)\Big\rangle\Big\|^2\Big] = \frac{1}{(1-\beta)m}\mathbb{E}\Big[\Big\|\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm te}\mathbf{X}^*, \mathbf{A}^{\rm te}\Big) - \Big\langle\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm te}\mathbf{x}, \mathbf{a}^{\rm te}\Big)\Big\rangle\Big\|^2\Big]. \quad (36)$$

Here the Gibbs bracket $\langle - \rangle$ is associated to the joint posterior (4) but with Δ replaced by Δ_μ in order to take into account that the noise varies in the training and test sets. Moreover $\mathbf{a}^{\rm te}$ is the restriction of $\mathbf{a}$ to its components belonging to the test set, namely $\mathbf{a}^{\rm te} = (a_\mu)_{\mu=\beta m+1}^m$. As all test samples are statistically equivalent, (36) can also be written as

$$\mathcal{E}_{\rm gen} = \mathbb{E}\Big[\Big(C_{\rm new} - \Big\langle\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}_{\rm new}\cdot\mathbf{x}, a\Big)\Big\rangle\Big)^2\Big] \quad (37)$$

where $(\mathbf{\Phi}_{\rm new}, C_{\rm new} = \varphi(\mathbf{\Phi}_{\rm new}\cdot\mathbf{X}^*/\sqrt{n}, A_{\rm new}))$ is a new, previously unobserved by the student, couple of pattern-label used by the teacher to test the student.

The generalization error quantifies the expected squared error between the labels of the test set and the *Bayes-optimal* estimator $\langle\varphi(\mathbf{\Phi}^{\rm te}\mathbf{x}, \mathbf{a}^{\rm te})\rangle$. This estimator is optimal in the sense that it minimizes the label mean-square-error among all possible student estimators. Note that in the general case where $\varphi(x, A)$ is stochastic in the sense that it depends on the random stream $\mathbf{A}$ (in contrast for example with the case of the binary perceptron $\varphi(x) = {\rm sgn}(x)$ that we will consider in the next section), the student has to learn *both* the $\mathbf{A}$ and $\mathbf{X}^*$ generated by the teacher in order to then be able to generalize.

We define the *optimal generalization error* as the limit $\lim_{\Delta^{\rm te}\to\infty}\mathcal{E}_{\rm gen}$. Our second main analytical result is its thermodynamic limit. We show in the next section that it is given by the following elegant formula (see Corollary 2.3 or (147) in Appendix J for the formula at finite $\Delta^{\rm te}$)

$$\lim_{\Delta^{\rm te}\to\infty}\lim_{n\to\infty}\mathcal{E}_{\rm gen} = \mathbb{E}_{V,a}\big[\varphi(\sqrt{\rho}\,V, a)^2\big] - \mathbb{E}_V\big[\mathbb{E}_{w,a}\big[\varphi(\sqrt{q^*}\,V + \sqrt{\rho - q^*}\,w, a)\big]^2\big], \quad (38)$$

where $V, w \overset{\rm iid}{\sim} \mathcal{N}(0,1)$ and $a \sim P_A$. In this expression q^* is a maximizer of the replica formula (18) evaluated at $\Delta^{\rm tr}$ (again, this takes into account that information about the model has been gained by the student during the learning). Note that as it should when $\Delta^{\rm te}\to\infty$, the optimal generalization error does not depend on the size of the test set.

The optimal generalization error should be independent of the random function representation as long as both φ and P_A are given to the student as it is the case in the present setting. The identity $\mathbb{E}_X\int dY\,Y^k P_{\rm out}^{\rm tr}(Y|X) = \mathbb{E}_{X,a}\int dY\,Y^k\exp\{-(Y - \varphi(X,a))^2/(2\Delta^{\rm tr})\}/\sqrt{2\pi\Delta^{\rm tr}}$ implies that the error (38) can be re-expressed equivalently as a function of the two first moments of the transition probability $P_{\rm out}^{\rm tr}$, that is

$$\lim_{\Delta^{\rm te}\to\infty}\lim_{n\to\infty}\mathcal{E}_{\rm gen} = -\Delta^{\rm tr} + \mathbb{E}_V\int dY\,Y^2 P_{\rm out}^{\rm tr}(Y|\sqrt{\rho}\,V) - \mathbb{E}_V\Big[\mathbb{E}_w\Big[\int dY\,Y P_{\rm out}^{\rm tr}(Y|\sqrt{q^*}\,V + \sqrt{\rho - q^*}\,w)\Big]^2\Big]. \quad (39)$$

Before showing how to compute the optimal generalization error, let us briefly discuss its "algorithmic" meaning. We assume that the student has access to unlimited computational power and can thus properly sample the posterior (4). Thus, as a proper Bayesian statistician, he samples a large amount $K \gg 1$ of pairs $\{(\mathbf{x}_i, \mathbf{a}_i^{\rm te})\}_{i=1}^K$ drawn according to the posterior (4) (e.g., by Monte Carlo sampling). Then for each such pair he computes the associated estimated vector of labels $\varphi(\mathbf{\Phi}^{\rm te}\mathbf{x}_i, \mathbf{a}_i^{\rm te})$ and performs the (componentwise) empirical average $K^{-1}\sum_{i=1}^K\varphi(\mathbf{\Phi}^{\rm te}\mathbf{x}_i, \mathbf{a}_i^{\rm te})$. This average converges to $\langle\varphi(\mathbf{\Phi}^{\rm te}\mathbf{x}, \mathbf{a}^{\rm te})\rangle$ as $K\to\infty$ which is the optimal student estimate of the labels of the test set.

We provide explicit formulas of the optimal generalization error for concrete applications in Sec. IV.

C. Computing the optimal generalization error

By a straightforward extension of the interpolation method presented in Sec. V one can generalize Theorem 2.1 to take into account that the noise variance Δ_μ differs in the training and test sets. This leads to an asymptotic free entropy rigorously given by

$$f_\infty = \sup_{q\in[0,\rho]}\inf_{r\geq 0} f_{\rm RS}^{\rm ts}(q,r) = \sup_{q\in[0,\rho]}\inf_{r\geq 0}\Big\{\psi_{P_0}(r) + \alpha\beta\Psi_{P_{\rm out}^{\rm tr}}(q;\rho) + \alpha(1-\beta)\Psi_{P_{\rm out}^{\rm te}}(q;\rho) - \frac{rq}{2}\Big\}. \quad (40)$$

Here $\Psi_{P_{\rm out}^{\rm tr}}$ and $\Psi_{P_{\rm out}^{\rm te}}$ are obtained from (16) using

$$P_{\rm out}\Big(Y_\mu\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu\Big) = \frac{1}{\sqrt{2\pi\Delta}}\int dP_A(a_\mu)e^{-\frac{1}{2\Delta}\big(Y_\mu - \varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu, a_\mu)\big)^2}, \quad (41)$$

We may also call this the "information theoretical generalization error" or "Bayes-optimal generalization error".

with $\Delta = \Delta^{\rm tr}$ and $\Delta = \Delta^{\rm te}$ respectively. Note that in the limit $\Delta^{\rm te} \to \infty$ (which means that the student has no access to the test set labels), $f_{\rm RS}^{\rm ts}$ (inside the brackets in (40)) collapses on $f_{\rm RS}$ given by (17) up to a trivial additive constant which corresponds to $\Psi_{P_{\rm out}^{\rm te}}$ in the high $\Delta^{\rm te}$ limit and a rescaling of the measurement rate $\alpha' = \alpha\beta$. This happens equivalently by taking $\beta = 1$ in $f_{\rm RS}^{\rm ts}$, which again means no information given to the student about the test set labels.

In order for our theorem to apply we consider the patterns $\mathbf{\Phi}_\mu$ to be made of i.i.d $\mathcal{N}(0,1)$ entries but we believe that the phenomenology that we present in this section applies to other models.

Let us now explain how to access the generalization error from the mutual information using I-MMSE relations. Denote the transition kernel before marginalization over a_μ

$$P_{\rm out}^{\rm te}\Big(Y_\mu\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu, a_\mu\Big) = \frac{1}{\sqrt{2\pi\Delta^{\rm te}}} e^{-\frac{1}{2\Delta^{\rm te}}\big(Y_\mu - \varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_\mu, a_\mu)\big)^2} . \tag{42}$$

and similarly for $P_{\rm out}^{\rm tr}$ where $\Delta^{\rm te}$ is replaced by $\Delta^{\rm tr}$. The free entropy f_n is nothing else than minus the Shannon (conditional) entropy density $-H(\mathbf{Y}|\mathbf{\Phi})/n$ of the distribution of the noisy labels. Thus it is related to the mutual information $I(\mathbf{X}^*, \mathbf{A}; \mathbf{Y}|\mathbf{\Phi}) = H(\mathbf{Y}|\mathbf{\Phi}) - H(\mathbf{Y}|\mathbf{X}^*, \mathbf{A}, \mathbf{\Phi})$ between the noisy labels and the classifer and stream $\mathbf{A}$ through

$$\begin{aligned}
i_n &:= \frac{1}{n} I(\mathbf{X}^*, \mathbf{A}; \mathbf{Y}|\mathbf{\Phi}) \\
&= -f_n - \frac{1}{n} H(\mathbf{Y}|\mathbf{X}^*, \mathbf{A}, \mathbf{\Phi}) \\
&= -f_n + \alpha\beta\, \mathbb{E}\int dY_1 P_{\rm out}^{\rm tr}\Big(Y_1\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1, A_1\Big) \ln P_{\rm out}^{\rm tr}\Big(Y_1\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1, A_1\Big) \\
&\qquad + \alpha(1-\beta)\mathbb{E}\int dY_m P_{\rm out}^{\rm te}\Big(Y_m\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_m, A_m\Big) \ln P_{\rm out}^{\rm te}\Big(Y_m\Big|\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_m, A_m\Big) \\
&= -f_n - \alpha\beta h_{\rm out}(\Delta^{\rm tr}) - \alpha(1-\beta) h_{\rm out}(\Delta^{\rm te}) ,
\end{aligned} \tag{43}$$

where

$$h_{\rm out}(\Delta) := -\mathbb{E}\int dY P_{\rm out}(Y|Z\sqrt{\rho}, A) \ln P_{\rm out}(Y|Z\sqrt{\rho}, A) = \frac{1}{2}(1 + \ln(2\pi\Delta)) \tag{44}$$

is the Shannon entropy of (42), that is of a Gaussian channel (here $A \sim P_A$, $Z \sim \mathcal{N}(0,1)$ and $\rho = \mathbb{E}[(X^*)^2]$). For the second equality in (43) we used that, conditionned on $\mathbf{\Phi}$ and $\mathbf{X}^*$, the $\{Y_\mu\}_{\mu=1}^{\beta m}$ are i.i.d as well as the $\{Y_\mu\}_{\mu=\beta m+1}^{m}$. For the third equality we used that conditionally on $\mathbf{X}^*$, $\{[\mathbf{\Phi X}^*]_\mu/\sqrt{n}\}_{\mu=1}^m$ are equal in distribution to i.i.d $\mathcal{N}(0,\rho)$ random variables.

Now in order to access the generalization error we employ the classical I-MMSE relation [63] for Gaussian noise but *restricted to the test set*. It takes the following form in the present setting (see Appendix A in [48] for a proof)

$$\frac{di_n}{d(\Delta^{\rm te})^{-1}} = \frac{(1-\beta)m}{2n}\mathcal{E}_{\rm gen} . \tag{45}$$

Here the arguments are the same as for the proof of Corollary 2.3, but we repeat them for the reader. Again we use that i_n is a sequence of concave functions in $(\Delta^{\rm te})^{-1}$ (see [48] for example) and thus its limit (which exists by Theorem 2.1) is concave too. Standard properties of convex sequences imply that $di_n/d(\Delta^{\rm te})^{-1}$ converges to $d\lim_{n\to\infty} i_n/d(\Delta^{\rm te})^{-1}$ at every $(\Delta^{\rm te})^{-1}$ at which $\lim_{n\to\infty} i_n$ is differentiable (which corresponds, by concavity, to $\mathbb{R}_+^*$ minus a countable subset). Therefore from (43), (44), (45) we find

$$\lim_{n\to\infty}\mathcal{E}_{\rm gen} = \Delta^{\rm te} - \frac{2}{\alpha(1-\beta)}\frac{d}{d(\Delta^{\rm te})^{-1}} \sup_{q\in[0,\rho]}\inf_{r\geq 0} f_{\rm RS}^{\rm ts}(q,r) . \tag{46}$$

It remains to evaluate this derivative using (40) to finally assess the asymptotic generalization error. A computation that we defer to Appendix J leads from (46) the formula (38) in the limit $\Delta^{\rm te} \to \infty$.

Let us mention that following the very same steps, one can also access the learning error defined as

$$\mathcal{E}_{\rm lea} := \frac{1}{\beta m}\mathbb{E}\Big[\Big\|\mathbf{C}^{\rm tr} - \Big\langle \varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm tr}\mathbf{x}, \mathbf{a}^{\rm tr}\Big)\Big\rangle\Big\|^2\Big] = \frac{1}{\beta m}\mathbb{E}\Big[\Big\|\varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm tr}\mathbf{X}^*, \mathbf{A}^{\rm tr}\Big) - \Big\langle \varphi\Big(\frac{1}{\sqrt{n}}\mathbf{\Phi}^{\rm tr}\mathbf{x}, \mathbf{a}^{\rm tr}\Big)\Big\rangle\Big\|^2\Big] \tag{47}$$

where the labels $\mathbf{C}$, the patterns $\mathbf{\Phi}$ and the stream $\mathbf{A}$ are now restricted to the training set. Note that this error vanishes in the limit $\Delta^{\rm tr} \to 0$ and this wathever the size of the training set. We also remark that a low training error does *not* imply a low generalization error. When the training set is small ($\alpha \ll 1$) the problem is underconstrained which prevents the student to reach a decent generalization error. The learning error can be obtained from the mutual information through

$$\frac{d}{d(\Delta^{\rm tr})^{-1}}\lim_{n\to\infty} i_n = \frac{\alpha\beta}{2}\lim_{n\to\infty}\mathcal{E}_{\rm lea} , \tag{48}$$

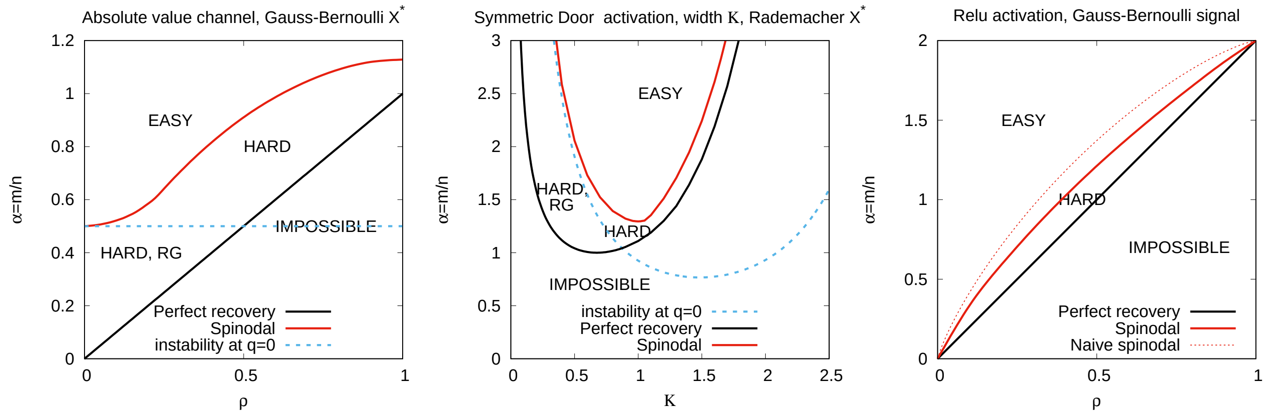

Fig. 1. Different phase diagrams showing the region where a perfect recovery is possible (these are noiseless problems). **Left:** The phase diagram for the absolute value problem with $\varphi(x) = |x|$ with a Gauss-bernoulli signal $P_0(x) = \rho \exp(-x^2/2)/\sqrt{2\pi} + (1-\rho)\delta(x)$, as a function of $\alpha = m/n$ and the fraction of non-zero components ρ. We find that a perfect recovery is *impossible* for $\alpha < \rho$. Perfect recovery becomes possible starting from $\alpha > \rho$, as in compressed sensing, but the problem seems numerically much harder. GAMP is not able to perform better than a random guess as long as $\alpha < 0.5$: We denote this region HARD, RG, for "not better than random guess". For larger values, the inference using GAMP leads better results than a purely random guess but cannot reach perfect recovery, so the problem remains HARD. GAMP can perfectly identify the hidden signal only for values of α larger than the so-called *spinodal* (or algorithmic threshold), when the problem becomes EASY. **Middle:** Phase diagram for the door function problem, with $\varphi(x) = 1$ if $-K < x < K$ and -1 else, for Rademacher signal $P_0(x) = (\delta(x-1) + \delta(x+1))/2$ as a function of α and K. The same regions and phenomenology are observed. **Right:** Phase diagram for the ReLU problem, with $\varphi(x) = \max(0, x)$, again with the EASY and HARD regions. Here it is always possible to perform better than chance using GAMP. The naive spinodal shows the algorithmic performance when using information only about the non-zero observations.

and is thus related to the potential (40) by

$$\lim_{n\to\infty} \mathcal{E}_{\rm lea} = \Delta^{\rm tr} - \frac{2}{\alpha\beta} \frac{d}{d(\Delta^{\rm tr})^{-1}} \sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}^{\rm ts}(q,r)\,. \tag{49}$$

The same computations as in Appendix J show that the final formula is given by the output error of Corollary 2.3 or, equivalently, the right hand side of (147) but with $\Delta^{\rm tr}$ replacing $\Delta^{\rm te}$ and evaluated at q^*, a maximizer of (18) (which we recall is the same value used in order to obtain the optimal generalization error).

IV. Application to concrete situations

In this section, we show how our results can be applied to many models of interest in fields ranging from machine learning to signal processing.

A. Optimal generalization error for some applications

Let us now give the explicit expression of the optimal generalization error for few relevant examples.

1) Sign channel, or perceptron: For the sign channal, the deterministic output (or "activation") function is $\varphi(x) = \mathrm{sgn}(x)$. This allows to model the so-called teacher-student perceptron problem in machine learning [43], or equivalently, the one-bit compressed sensing in signal processing [23]. Both situations have been discussed in details within the replica formalism (see for instance [45], [50], [53], [64]) and we confirm all these heuristic computations within our approach. Let $V \sim \mathcal{N}(0,1)$. The formula (38) for the generalization error then reduces to

$$\lim_{\Delta^{\rm te}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\rm gen} = 1 - \int dV \frac{e^{-\frac{V^2}{2}}}{\sqrt{2\pi}} \Big\{ \frac{2}{\sqrt{\pi}} \int_0^{V\sqrt{\frac{q^*}{2(\rho-q^*)}}} dt\, e^{-t^2} \Big\}^2 = 1 - \mathbb{E}\Big[\mathrm{erf}\Big(V\sqrt{\frac{q^*}{2(\rho-q^*)}}\Big)^2\Big]\,. \tag{50}$$

2) Linear regression: The additive white Gaussian noise (AWGN), or linear regression, is defined by $\varphi(x, A) = x + \sigma A$ with $A \sim \mathcal{N}(0,1)$. This models the (noisy) linear regression problem, as well as a noisy random linear estimation and compressed sening. In this case (38) leads

$$\lim_{\Delta^{\rm te}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\rm gen} = \rho - q^* + \sigma^2\,. \tag{51}$$

This result agrees with [7] in the limit $\sigma \to 0$.

3) Rectified linear unit (ReLU): Another example of deterministic output function is the ReLU where $\varphi(x) = \max(0, x)$. This channel models the behavior of a single neuron with the celebrated rectified linear unit activation [1] ubiquitous in multilayer neural networks. In this case (38) becomes after simple algebra and Gaussian integrations (again $V \sim \mathcal{N}(0,1)$),

$$\lim_{\Delta^{\rm te}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\rm gen} = \frac{\rho}{2} - \frac{q^*}{4}\Big(1 + \mathbb{E}_V\Big[\Big\{V \mathrm{erf}\Big(V\sqrt{\frac{q^*}{2(\rho-q^*)}}\Big)\Big\}^2\Big]\Big) - \frac{(\rho-q^*)^{3/2}}{\sqrt{\rho+q^*}}\Big(\frac{1}{2\pi} + \frac{q^*}{\rho\pi}\sqrt{\frac{\rho+q^*}{\rho-q^*}}\Big)\,. \tag{52}$$

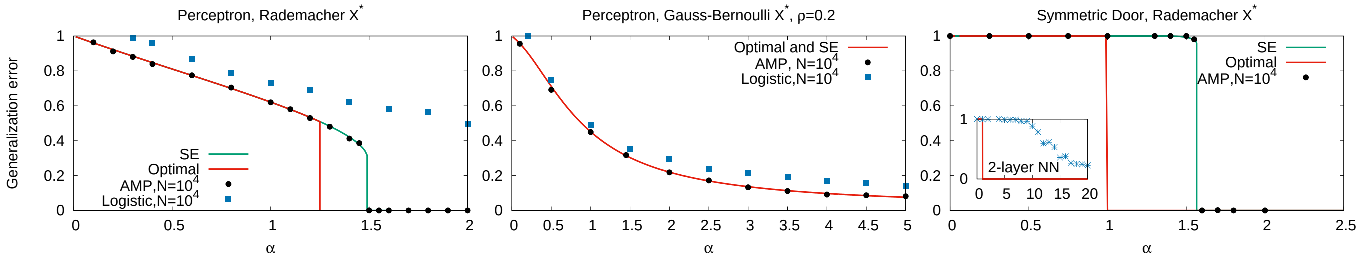

Fig. 2. Generalization error in three classification problems versus α, the size of the training set being αn. The red line is the Bayes-optimal generalization error ((50) for the perceptron or (56) for the symmetric door) while the green one shows the (asymptotic) performances of GAMP as predicted by the state evolution (SE) [62], when different. For comparison, we also show the result of GAMP (black dots) and, in blue, the performance of a standars out-of-the-box solver. **Left:** Perceptron, with $\varphi(x) = \mathrm{sgn}(x)$ and a Rademacher signal. While a perfect generalization is information theoretically possible starting from $\alpha = 1.249(1)$, the state evolution predicts that GAMP will allow such perfect prediction only from $\alpha = 1.493(1)$. The results of a logistic regression with fine-tuned ridge penalty with the software scikit-learn [65] are shown for comparison. **Middle:** Perceptron with Gauss-Bernoulli coefficients for the signal. No phase transition is observed in this case, but a smooth decrease of the error with α. The results of a logistic regression with fine-tuned ℓ_1 sparsity-enhancing penalty (again with [65]) are very close to optimal. **Right:** The symmetric door activation rule with parameter K chosen in order to observe the same number of occurence of the two classes. In this case there is a sharp phase transition at $\alpha = 1$ from a situation where it is impossible to learn the rule, so that the generalization is not better than a random guess, to a situation where the generalization error drops to zero. However, GAMP identifies the rule correctly only starting from $\alpha = 1.5$. Interestingly, this non linear rule seems very hard to learn. Using Keras [66], a neural network with 2 hidden layers was able to learn approximately the rule, but only for much larger training set sizes.

4) Sigmoid, or logistic regression: Let us also consider a stochastic output function. After having generated the classifier, the teacher randomly associates the label $+1$ to the pattern $\mathbf{\Phi}_\mu$ with probability $f_\lambda(\mathbf{\Phi}_\mu \cdot \mathbf{X}^*)$, where $f_\lambda(x) = (1+\exp(-\lambda x))^{-1} \in [0,1]$ is the sigmoid of parameter $\lambda > 0$, and the label -1 with probability $1 - f_\lambda(\mathbf{\Phi}_\mu \cdot \mathbf{X}^*)$. One of the (many) possible ways for the teacher to do so is by selecting $\varphi(x, A) = \mathbb{I}(A \le f_\lambda(x)) - \mathbb{I}(A > f_\lambda(x))$, where $\mathbb{I}(E)$ is the indicator function of the event E. He then generates a stream of uniform random numbers $\mathbf{A} \overset{\text{iid}}{\sim} \mathcal{U}_{[0,1]}$ and obtain the labels through (35). Let $V, w \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. In this setting the error (38) becomes

$$\lim_{\Delta^{\text{te}}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\text{gen}} = 2 - 4\,\mathbb{E}_V\left[\left\{\mathbb{E}_w f_\lambda(\sqrt{q^*}\,V + \sqrt{\rho - q^*}\,w)\right\}^2\right]. \tag{53}$$

This fomula reduces to (50) when $\lambda \to \infty$ as it should.

5) Absolute value: A further example of a purely deterministic output function is the absolute value where $\varphi(x) = |x|$. This models a situation similar to compressed sensing, except that the sign of the output has been lost. It could be seen as a simple version of the phase retrival problem. In this case (38) becomes

$$\lim_{\Delta^{\text{te}}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\text{gen}} = \rho - \mathbb{E}_V\left[b(V\sqrt{q^*}, \rho - q^*)^2\right], \tag{54}$$

where

$$b(x,y) = \sqrt{\frac{2y}{\pi}}e^{-\frac{x^2}{2y}} + \frac{x}{2}\mathrm{erfc}\Big(-\frac{x}{\sqrt{2y}}\Big) - \frac{x}{2}\Big\{1 + \mathrm{erf}\Big(-\frac{x}{\sqrt{2y}}\Big)\Big\}. \tag{55}$$

6) Symmetric Door: A final example of deterministic output function is the symmetric door where $\varphi(x) = 1$ if $-K < x < K$ and -1 otherwise. In this case (38) becomes

$$\lim_{\Delta^{\text{te}}\to\infty} \lim_{n\to\infty} \mathcal{E}_{\text{gen}} = 1 - \mathbb{E}_V\left[\left\{\mathrm{erf}\Big(\frac{K - \sqrt{q}\,V}{\sqrt{2(\rho - q)}}\Big) - \mathrm{erf}\Big(\frac{-K - \sqrt{q}\,V}{\sqrt{2(\rho - q)}}\Big) - 1\right\}^2\right]. \tag{56}$$

Again, many other situations can be directly considered, including stochastic ones, for instance the probit and logit regressions.

B. Phase diagrams: Easy, hard and impossible estimation and learning phases

First, we shall consider three deterministic channels and ask (we consider noiseless problems): How many measurements are needed in order to perfectly recover the signal? In the case of the linear channel, this question has been adressed in great details for the compressed sensing case [24], [54], and we find a simular phenomenology here, albeit with some subtelties.

1) The Relu channel: Let us start by discussing the case of the ReLU channel, with a signal coming from a Gauss-Bernoulli distribution with a fraction ρ of non-zero (Gaussian) values. In this case, our analysis shows that a perfect generalization (and thus a perfect reconstruction of the signal as well) is possible whenever $\alpha > 2\rho$, and impossible when $\alpha < 2\rho$. This is very intuitive, since half of the measurements (those non-zero) are giving as much information as in the linear case, thus the factor 2. How hard is it to actually solve the problem in practice? The answer is given by applying the state evolution analysis to GAMP, which tells us that only for even larger values of α, beyond the so-called spinodal transition, does GAMP reach a perfect recovery. Notice, however, that this spinodal transition occurs at a significantly lower measurement rate α than one

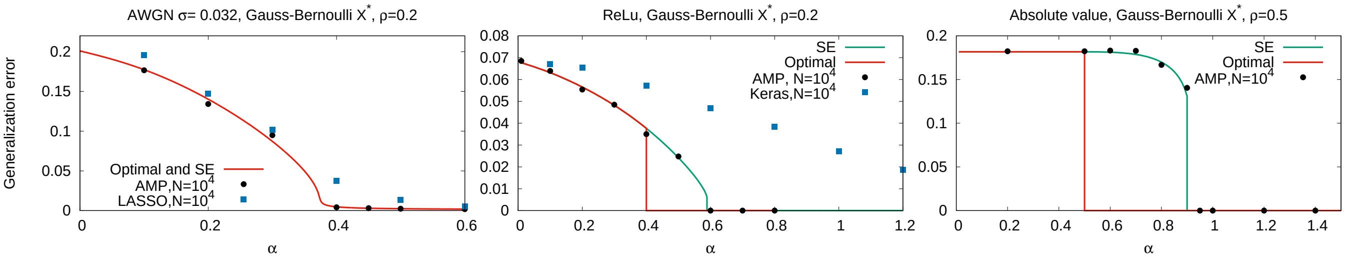

Fig. 3. Same as Fig. 2 but for regression problems. The generalization error is plotted as a function of α, the size of the training set being αn. The red line is again the Bayes-optimal generalization error ((51) for AWGN, (52) for the ReLU and (54) for the absolute value) while the green one shows the (asymptotic) performances of GAMP as predicted by the state evolution (SE) [62], when different. Again, we also show the result of GAMP on a particular instance (black dots) and, in blue, the performance of an out-of-the-box solver. **Left:** The first example is with an additive white Gaussian noise and a Gauss-Bernoulli signal. For this choice of noise, there is no sharp transition (as opposed to what happens at smaller noises). The results of a LASSO with fine-tuned ℓ_1 sparsity-enhancing penalty (with [65]) are very close to optimal. **Middle:** Here we used a ReLU-type function $\varphi(x) = \max(0, x)$, still with a Gauss-Bernoulli signal. Now there is a phase transition at $\alpha = 2\rho = 0.4$, but GAMP requires more samples to reach perfect recovery. We show for comparison the result of a maximum likelihood estimation performed with Keras [66]. **Right:** The last example shows the result for the absolute value function $\varphi(x) = |x|$. The perfect recovery starts at $\alpha = \rho = 0.5$, but the problem is again harder algorithmically for GAMP.

would reach just keeping the non-zero measurements. This shows that, actually, these zero measurements contains a useful information for the algorithm. The situation is shown in the right side of Fig. 1.

What we have discussed here is the appearance of a very generic scenario, namely the presence of two different transitions when trying to solve the Bayesian optimal problem: For $\alpha < 2\rho$, it is information theoretically impossible to identify perfectly the signal. We refer to this situation as the IMPOSSIBLE phase. For $\alpha > 2\rho$, reconstruction is theoretically possible, however, we do not know any polynomial-time algorithm that would succeed unless $\alpha > \alpha_{\rm spinodal}$, where GAMP provably finds the hidden assignment. So we further divide the POSSIBLE phase into the EASY and HARD regions.

2) The absolute value channel: We know move to the absolute value channel. We observe again a similar EASY, HARD and IMPOSSIBLE phases scenario. Here, the analysis of the mutual information shows that a perfect reconstruction is possible as soon as $\alpha > \rho$: In other words, the problem is –information theoretically– as easy, or as hard as the compressed sensing one. This is maybe less surprising when one think of the following algorithm: Try all 2^m choices of the possible signs for the m outputs, and solve a compressed sensing problem for each of them. Clearly, this should yields a perfect solution only in the case of the actual signal.

Algorithmically, however, the problem is much harder than before. As shown in the left side of Fig. 1, one requires a much larger fraction α of measurements in order for GAMP to actually solve the problem. Besides, another interesting phenomenon occurs (which is in fact a characteristic of symmetric $\varphi(x)$ functions): There is always an extremum of the mutual information with an overlap value $q = 0$. For this problem, this extremum is actually "stable" (meaning that it is actually a minimum in q) for all $\alpha < 0.5$. This has the two following implications: $i)$ In the IMPOSSIBLE phase, when $\alpha < 0.5$ and $\rho > \alpha$, the minimum at $q = 0$ is actually the global one. In this case, the MMSE and the generalization error are the ones given by using 0 as a guess for each element of $\mathbf{X}^*$; in other words, there is no useful information that one can exploit and no algorithmic approach can be better than a random guess! $ii)$ In the POSSIBLE but HARD phase when $\alpha < 0.5$, GAMP initialised at random, infinitely close to the $q = 0$ fixed point, will remain there. This suggests that in this region, even if a perfect reconstruction is possible, it will anyway be very hard to beat a random guess in practice. We thus further divide the HARD phase into the HARD and HARD, RG phase (where RG stands for random guess).

3) The symmetric door channel: We finally discuss a last situation that is the symmetric door channel, for a Rademacher signal $\mathbf{X}^*$ where each element is chosen at random between $+1$ and -1. In this case we find again, as in the absolute value problem, the EASY, HARD, HARD RG and IMPOSSIBLE phases.

C. *Generalization in classification problems*

We now discuss these results in the context of supervised classification (that is, a ± 1 output) in the teacher-student scenario. Again, we select three particular cases and illustrate our results in Fig. 2. For the purpose of the discussion, we consider two deterministic problems: The sign output (perceptron) and the symmetric door one. Within these examples, perhaps the most interesting one is the latter, where we use the symmetric door with $\kappa = 0.674489...$, a value chosen such that the output produces as many $+1$ than -1.

D. *Generalization in regression problems*

We finally discuss these results in the context of supervised regression problems (that is, problems where the output is real valued), in the teacher-student scenario. We select three particular cases, and illustrate our results in Fig. 3. We choose to consider one output function with randomness and two deterministic ones.

V. Proof of the replica formula by the stochastic interpolation method

We now prove Theorem 2.1. Our main tool will be an interpolation method recently introduced in [46] and called "stochastic interpolation method" (for reasons that will not be discussed here). Here we formulate the method as a direct evolution of the Guerra and Toninelli interpolation method developed in the context of spin glasses [67]. In constrast with the discrete and more pedestrian version of the stochastic interpolation method presented in [46], here we employ a continuous approach which is more straightforward (see [46] for the links between the discrete and continuous versions of the method).

We will prove Theorem 2.1 under the following hypotheses:

(H1) The prior distribution P_0 has a bounded support.

(H2) φ is a bounded $\mathcal{C}^2$ function with bounded first and second derivatives w.r.t its first argument.

These assumptions will then be relaxed in Appendix F to the assumptions (h1) and (h2). Since the observations (1) are equivalent to the rescaled observations

$$\widetilde{Y}_\mu := \Delta^{-1/2}\, Y_\mu = \Delta^{-1/2}\,\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu, A_\mu\Big) + Z_\mu\,, \qquad 1 \le \mu \le m\,, \tag{57}$$

the variance Δ of the Gaussian noise can be "incorporated" inside the function φ. Thus, it suffices to prove Theorem 2.1 for $\Delta = 1$ and we will now suppose, for the rest of the proof, to be in this equivalent case.

A. Interpolating estimation problem

We introduce an "interpolating estimation problem" that interpolates between the orginal problem (2) at $t = 0$, $t \in [0,1]$ being the interpolation parameter, and the two scalar problems described in Sec. II-B at $t = 1$ which are anaytically tractable. For $t \in]0,1[$ the interpolating estimation problems is a mixture of the original and scalar problems. This interpolation scheme is inspired from the interpolation paths used by Talagrand to study the perceptron, see [68]. But thanks to a novel ingredient specific to the stochastic interpolation method, it allows to obtain much stronger results, namely a complete proof of the replica formula instead of the bounds that are generally obtained by more classical interpolation methods.

Let $q : [0,1] \to [0,\rho]$ be a continuous "interpolation function" and $r \ge 0$. Define

$$S_{t,\mu} := \sqrt{\frac{1-t}{n}}\,[\mathbf{\Phi}\mathbf{X}^*]_\mu + \sqrt{\int_0^t q(v)dv}\; V_\mu + \sqrt{\int_0^t (\rho - q(v))dv}\; W^*_\mu \tag{58}$$

where $V_\mu, W^*_\mu \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Consider the following observation channels, with two types of observations obtained through

$$\begin{cases} Y_{t,\mu} & \sim \quad P_{\rm out}(\,\cdot \,|\, S_{t,\mu})\,, \quad \text{for } 1 \le \mu \le m, \\ Y'_{t,i} & = \quad \sqrt{r\,t}\, X^*_i + Z'_i\,, \quad \text{for } 1 \le i \le n, \end{cases} \tag{59}$$

where $Z'_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We assume that $\mathbf{V} = (V_\mu)_{\mu=1}^m$ is known and that the inference problem is to recover both $\mathbf{W}^* = (W^*_\mu)_{\mu=1}^m$ and $\mathbf{X}^* = (X^*_i)_{i=1}^n$ from the "time-dependent" observations $\mathbf{Y}_t = (Y_{t,\mu})_{\mu=1}^m$ and $\mathbf{Y}'_t = (Y'_{t,i})_{i=1}^n$.

We now understand that rt appearing in the second set of measurements in (59), and the terms $1-t$, $\int_0^t q(v)dv$ and $\int_0^t(\rho - q(v))dv$ appearing in the first set all play the role of signal-to-noise ratios in the interpolating model, with t giving more and more "power" (or weight) to the scalar inference channels when increasing. Here is the first crucial and novel ingredient of our interpolation scheme. In the classical interpolation method, these snr would all take a trivial form (i.e would be linear in t) but here, the non-trivial (integral) dependency in t of the two latter snr through the use of the interpolation function q allows for much more flexibility when choosing the interpolation path. This will allow us to actually choose the "optimal interpolation path" (this will become clear soon).

Define $u_y(x) := \ln P_{\rm out}(y|x)$ and, with a slight abuse of notations,

$$s_{t,\mu} = s_{t,\mu}(\mathbf{x}, w_\mu) := \sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{x}]_\mu + \sqrt{\int_0^t q(v)dv}\; V_\mu + \sqrt{\int_0^t (\rho - q(v))dv}\; w_\mu\,. \tag{60}$$

We introduce the *interpolating Hamiltonian*

$$\mathcal{H}_t(\mathbf{x}, \mathbf{w}; \mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi}) := -\sum_{\mu=1}^m \ln P_{\rm out}(Y_{t,\mu}|s_{t,\mu}) + \frac{1}{2}\sum_{i=1}^n \big(Y'_{t,i} - \sqrt{t\,r}\, x_i\big)^2 \tag{61}$$

$$= -\sum_{\mu=1}^m u_{Y_{t,\mu}}(s_{t,\mu}) + \frac{1}{2}\sum_{i=1}^n \big(\sqrt{t\,r}\,(X^*_i - x_i) + Z'_i\big)^2\,, \tag{62}$$

and the corresponding Gibbs bracket $\langle - \rangle_t$ which is the expectation operator w.r.t the t-dependent posterior distribution of $(\mathbf{x}, \mathbf{w})$ given the observations $(\mathbf{Y}_t, \mathbf{Y}'_t)$. It is defined as

$$\langle g(\mathbf{x}, \mathbf{w}) \rangle_t := \frac{1}{\mathcal{Z}_t(\mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi})} \int dP_0(\mathbf{x}) \mathcal{D}\mathbf{w}\, g(\mathbf{x}, \mathbf{w})\, e^{-\mathcal{H}_t(\mathbf{x}, \mathbf{w}; \mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi})}\,, \tag{63}$$

for every continuous bounded function g on $\mathbb{R}^n \times \mathbb{R}^m$. In (63) $\mathcal{D}\mathbf{w} = (2\pi)^{-m/2}\prod_{\mu=1}^{m} dw_\mu e^{-w_\mu^2/2}$ is the m-dimensional standard Gaussian and $\mathcal{Z}_t(\mathbf{Y}_t, \mathbf{Y}'_t, \mathbf{\Phi})$ is the appropriate normalization:

$$\mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{\Phi}) := \int dP_0(\mathbf{x}) D\mathbf{w}\, e^{-\mathcal{H}_t(\mathbf{x},\mathbf{w};\mathbf{Y},\mathbf{Y}')}\,. \tag{64}$$

Finally the *interpolating free entropy* is

$$f_n(t) := \frac{1}{n}\mathbb{E}\ln \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{\Phi}) = \mathbb{E}_{\mathbf{\Phi}}\int d\mathbf{Y} d\mathbf{Y}' \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{\Phi}) \ln \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}', \mathbf{\Phi})\,. \tag{65}$$

One verifies easily that

$$\begin{cases} f_n(0) &= f_n - 1/2\,, \\ f_n(1) &= \psi_{P_0}(r) - \frac{1+r\rho}{2} + \frac{m}{n}\Psi_{P_{\rm out}}(\int_0^1 q(t)dt; \rho)\,. \end{cases} \tag{66}$$

Here is really another crucial property of the interpolating model that we emphasize: It is such that at $t = 0$ we recover the original model and thus $f_n(0) = f_n - 1/2$ (the trivial constant comes from the purely noisy measurements of the second channel in (59)), while at $t = 1$ we have the two scalar inference channels and thus the associated terms ψ_{P_0} and $\Psi_{P_{\rm out}}$ appear in $f_n(1)$. These are precisely the terms appearing in the potential (17). This is the reason for the introduction of these scalar channels.

B. Free entropy variation along the interpolation path

From the understanding of the previous section, it is at this stage very natual to evaluate the variation of free entropy along the interpolation path, which allows to "compare" the original and purely scalar models thanks to the identity

$$f_n = f_n(0) + \frac{1}{2} = f_n(1) - \int_0^1 \frac{df_n(t)}{dt} + \frac{1}{2}\,, \tag{67}$$

where the first equality follows from (66). As discussed above, part of the potential (17) appears in $f_n(1)$. If the interpolation is properly done, the missing terms required to obtain the potential on the r.h.s of (67) should naturally appear. Then by choosing the optimal interpolation path thanks to the non-trivial snr dependencies in t (i.e by selecting the proper interpolating function q), we will be able to show the equality between the replica formula and the true entropy f_n.

We thus now compute the t-derivative of the free entropy along the interpolation path (see Appendix D for the proof).

***Proposition** 5.1 (Free entropy variation):* For model (2), the t-derivative of the free entropy (65) verifies

$$\frac{df_n(t)}{dt} = -\frac{1}{2}\mathbb{E}\Big\langle\Big(\frac{1}{n}\sum_{\mu=1}^{m} u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu}) - r\Big)\big(Q - q(t)\big)\Big\rangle_t + \frac{rq(t)}{2} - \frac{r\rho}{2} + \mathcal{O}_n(1), \tag{68}$$

where $\mathcal{O}_n(1)$ is a quantity that goes to 0 in the $n, m \to \infty$ limit, uniformly in $t \in [0,1]$ and the *overlap* is

$$Q := \frac{1}{n}\sum_{i=1}^{n} X_i^* x_i\,. \tag{69}$$

C. Overlap concentration

The next lemma plays a key role in our proof. Essentially it states that the overlap concentrates arounds its mean, a behavior called "replica symmetric" in statistical physics. Similar results have been obtained in the context of the analysis of spin glasses [68]. Here we use a formulation taylored to Bayesian inference problems as developed in the context of LDPC codes, random linear estimation [48] and Nishimori symmetric spin glasses [69]–[71].

We introduce a "small" perturbation of the interpolating estimation problem by *adding* to the Hamiltonian (62) a term

$$\sum_{i=1}^{n}\Big(\epsilon\frac{x_i^2}{2} - \epsilon x_i X_i^* - \sqrt{\epsilon}x_i\widehat{Z}_i\Big) \tag{70}$$

where $(\widehat{Z}_i)_{i=1}^n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. This term can be interpreted as having a set of extra observations coming from a Gaussian side-channel $\widehat{Y}_i = \sqrt{\epsilon}X_i^* + \widehat{Z}_i$ and preserves the Nishimori identity (see Appendix A). The new Hamiltonian $\mathcal{H}_{t,\epsilon}(\mathbf{x}, \mathbf{w}; \mathbf{Y}, \mathbf{Y}', \mathbf{\Phi})$ defines a new Gibbs bracket $\langle - \rangle_{n,t,\epsilon}$ and free entropy $f_{n,\epsilon}(t)$, and all the set up of Sec. V-A and Proposition 5.1 trivially extend. This perturbation induces only a small change in the free entropy, namely of the order of ϵ:

***Lemma** 5.2 (Small free entropy variation under perturbation):* Let $C_0 > 0$ such that the support of P_0 is included in $[-C_0, C_0]$. For all $\epsilon > 0$ and all $t \in [0,1]$,

$$|f_{n,\epsilon}(t) - f_n(t)| \le \epsilon\frac{C_0^2}{2} \tag{71}$$

Proof: A simple computation gives

$$\Big|\frac{\partial f_{n,\epsilon}(t)}{\partial \epsilon}\Big| = \frac{1}{2}\big|\mathbb{E}\langle Q\rangle_{n,t,\epsilon}\big| \le \frac{C_0^2}{2}\,, \tag{72}$$

which proves the lemma. ■

Moreover, this small perturbation forces the overlap to concentrates around their mean:

***Lemma** 5.3 (Overlap concentration):* For any $0 < a < b < 1$,

$$\lim_{n\to\infty}\int_a^b d\epsilon \int_0^1 dt\, \mathbb{E}\big\langle \big(Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon}\big)^2\big\rangle_{n,t,\epsilon} = 0\,. \tag{73}$$

In Appendix I we briefly sketch the main steps of the proof for the convenience of the reader and refer to [46] for more details where the method has been streamlined.

Lemma 5.3 implies that there exists a sequence $(\epsilon_n)_{n\ge1} \in (0,1)^{\mathbb{N}^*}$ that converges to 0 such that

$$\lim_{n\to\infty}\int_0^1 dt\, \mathbb{E}\big\langle \big(Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon_n}\big)^2\big\rangle_{n,t,\epsilon_n} = 0\,. \tag{74}$$

$(\epsilon_n)_{n\ge1}$ converges to 0, so Lemma 5.2 gives that $f_{n,\epsilon_n}(t)$ and $f_n(t)$ have the same limit (provided it exists). In the rest of the section, in order to alleviate the notations, we abusively remove the perturbation subscript ϵ_n since it makes no difference for the computation of the limit of the free entropy.

D. Cancelling the remainder

Note from (66) and (17) that the second and third terms appearing in (68) are precisely the missing ones that are required in order to obtain the expression of the potential on the r.h.s of (67). Thus in order to prove Theorem 2.1 we would like to "cancel" the Gibbs bracket in (68), which is the so called *remainder* (once integrated over t). This is made possible thanks to the new ingredients specific to the stochastic interpolation method. To do so, we would like to choose $q(t) = \mathbb{E}\left\langle Q\right\rangle_t$, which is approximately equal to Q because it concentrates, see Lemma 5.3. However, $\mathbb{E}\left\langle Q\right\rangle_t$ depends on $\int_0^t q(v)dv$. The equation $q(t) = \mathbb{E}\left\langle Q\right\rangle_t$ is therefore an order 1 differential equation over q, written in integral form.

***Proposition** 5.4 (Existence of the optimal interpolation function):* For all $r \ge 0$ the differential equation

$$q(t) = \mathbb{E}\left\langle Q\right\rangle_t \tag{75}$$

admits a unique solution $q_n^{(r)}$ on $[0,\rho]$ and the mapping

$$r \ge 0 \mapsto \int_0^1 q_n^{(r)}(v)dv \tag{76}$$

is continuous.

Proof: One verify easily that $\mathbb{E}\left\langle Q\right\rangle_t$ is a bounded $\mathcal{C}^1$ function of $(\int_0^t q(v)dv, r)$. The proposition follows then from an application of the parametric Cauchy-Lipschitz theorem. ■

Using this optimal choice for the interpolating function allows then to relate the potential and free entropy.

***Proposition** 5.5 (Linking free entropy and potential):* Let $(r_n)_{n\ge1} \in \mathbb{R}_+^{\mathbb{N}}$ be a bounded sequence. For $n \in \mathbb{N}$, let $q_n^{(r_n)}$ be the solution of (75). Then

$$f_n = f_{\rm RS}\Big(\int_0^1 q_n^{(r_n)}(v)dv, r_n\Big) + \mathcal{O}_n(1)\,. \tag{77}$$

Proof: $q_n^{(r_n)}$ satisfies (75). Therefore by the Cauchy-Schwarz inequality

$$\begin{aligned}&\Big|\int_0^1 dt\, \mathbb{E}\Big\langle\Big(\frac{1}{n}\sum_{\mu=1}^m u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu}) - r_n\Big)\big(Q - q_n^{(r_n)}(t)\big)\Big\rangle_t\Big| \\ &\qquad \le \Big(\int_0^1 dt\, \mathbb{E}\Big\langle\Big(\frac{1}{n}\sum_{\mu=1}^m u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu}) - r_n\Big)^2\Big\rangle_t\Big)^{1/2}\Big(\int_0^1 dt\, \mathbb{E}\big\langle\big(Q - q_n^{(r_n)}(t)\big)^2\big\rangle_t\Big)^{1/2} = \mathcal{O}_n(1)\,.\end{aligned} \tag{78}$$

The last equality uses that the first factor is bounded (independently of t) because we supposed that $P_{\rm out}$ is generated by (57) with assumptions (H1) and (H2) (see Appendix E for proof details) and the second factor goes to 0 when $n, m \to \infty$ by (74), (75). Therefore from (68)

$$\int_0^1 \frac{df_n(t)}{dt} dt = \frac{r_n}{2} \int_0^1 q_n^{(r_n)}(t) dt - \frac{r_n \rho}{2} + \mathcal{O}_n(1) \,. \tag{79}$$

When plugging this identity in (67) and combining this with (66) we reach

$$f_n = \psi_{P_0}(r_n) + \frac{m}{n} \Psi_{P_{\rm out}} \Big(\int_0^1 q_n^{(r_n)}(t) dt; \rho \Big) - \frac{r_n}{2} \int_0^1 q_n^{(r_n)}(t) dt + \mathcal{O}_n(1) \,. \tag{80}$$

Recalling that $m/n \to \alpha$ as $m, n \to \infty$ allows to recognize from (17) the claimed identity (77). ■

E. Lower and upper matching bounds

We now possess all the necessary tools to prove Theorem 2.1 in two steps. $i)$ We start by proving that $\lim_{n\to\infty} f_n = \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r)$. Recall that for the moment we assume the stronger hypotheses (H1) and (H2). $ii)$ Once this is done we can prove that moreover $\lim_{n\to\infty} f_n = \sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}(q,r)$ using the following arguments. $ii_a)$ Under hypotheses (H1), (H2) the Corollary G.2 of Appendix G applies and allows to assert $\sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) = \sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}(q,r)$. $ii_b)$ This combined with $\lim_{n\to\infty} f_n = \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r)$ proven in step $i)$ leads that $\lim_{n\to\infty} f_n = \sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}(q,r)$ under (H1), (H2). $ii_c)$ Finally, the approximation arguments given in Appendix F permit to relax (H1), (H2) to the weaker hypotheses (h1), (h2) and thus to obtain the second (from step $ii)$) equality of Theorem 2.1.

We defer to Appendix G the proof of the last equality, namely that this "$\sup \inf$" is attained at the supremum of the state evolution fixed points, see Lemma G.4.

We now tackle step $i)$. Let us start by the lower bound.

***Proposition** 5.6 (Lower bound):* The free entropy (10) verifies

$$\liminf_{n\to\infty} f_n \geq \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) \,. \tag{81}$$

Proof: By Proposition 5.5 we have that for any $r \geq 0$,

$$f_n \geq f_{\rm RS}\Big(\int_0^1 q_n^{(r)}(t) dt, r \Big) + \mathcal{O}_n(1) \geq \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) + \mathcal{O}_n(1) \tag{82}$$

and thus

$$\liminf_{n\to\infty} f_n \geq \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) \,. \tag{83}$$

This is true for all $r \geq 0$ thus we obtain Proposition 5.6. ■

We now turn our attention to the converse bound.

***Proposition** 5.7 (Upper bound):* The free entropy (10) verifies

$$\limsup_{n\to\infty} f_n \leq \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) \,. \tag{84}$$

Proof: Let $K = 2\alpha \Psi'_{P_{\rm out}}(\rho;\rho)$. The mapping from equation (76) is continuous, consequently the application

$$\begin{array}{ccc} [0,K] & \to & [0,K] \\ r & \mapsto & 2\alpha \Psi'_{P_{\rm out}}\big(\int_0^1 q_n^{(r)}(t) dt; \rho \big) \end{array} \tag{85}$$

is continuous (recall that $\Psi'_{P_{\rm out}}$ denotes the derivative of $\Psi_{P_{\rm out}}$ w.r.t its first argument, and is shown to be continuous and bounded in Appendix B). It admits therefore a fixed point $r_n^* = 2\alpha \Psi'_{P_{\rm out}}(\int_0^1 q_n^{(r_n^*)}(t) dt; \rho)$. Proposition 5.5 gives then

$$f_n = f_{\rm RS}\Big(\int_0^1 q_n^{(r_n^*)}(t) dt, r_n^* \Big) + \mathcal{O}_n(1) \,. \tag{86}$$

We now remark that

$$f_{\rm RS}\Big(\int_0^1 q_n^{(r_n^*)}(t) dt, r_n^* \Big) = \inf_{q\in[0,\rho]} f_{\rm RS}(q, r_n^*) \,. \tag{87}$$

Indeed, the function $g_{r_n^*} : q \in [0,\rho] \mapsto f_{\rm RS}(q, r_n^*)$ is convex (because of Lemma B.1) and its derivative is

$$g'_{r_n^*}(q) = \alpha \Psi'_{P_{\rm out}}(q) - \frac{r_n^*}{2}\,. \tag{88}$$

Since $g'_{r_n^*}(\int_0^1 q_n^{(r_n^*)}(t)dt) = 0$ by definition of r_n^*, the minimum of $g_{r_n^*}$ is necessarily achieved at $\int_0^1 q_n^{(r_n^*)}(t)dt$. Combining (86) with (87) we reach

$$f_n = \inf_{q\in[0,\rho]} f_{\rm RS}(q, r_n^*) + \mathcal{O}_n(1) \leq \sup_{r\geq 0} \inf_{q\in[0,\rho]} f_{\rm RS}(q,r) + \mathcal{O}_n(1) \tag{89}$$

which allows to deduce Proposition 5.7. ■

From the arguments given at the beginning of the section, this ends the proof of Theorem 2.1.

Appendix A
The Nishimori identity

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

$$\mathbb{E}\langle f(\mathbf{Y},\mathbf{X}^{(1)},\dots,\mathbf{X}^{(k)})\rangle = \mathbb{E}\langle f(\mathbf{Y},\mathbf{X}^{(1)},\dots,\mathbf{X}^{(k-1)},\mathbf{X})\rangle\,. \tag{90}$$

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

Appendix B
Some properties of the scalar channel

We prove here some properties of the free entropy of the second scalar channel (15). In this section, we will keep the dependence in ρ of $\Psi_{P_{\rm out}}(q,\rho)$ implicit, and write simply $\Psi_{P_{\rm out}}(q)$. The derivatives of this function are taken w.r.t q.

Let

$$\begin{array}{cccc} P_{\rm out}: & \mathbb{R}^2 & \to & \mathbb{R}_+ \\ & (x,y) & \mapsto & P_{\rm out}(y|x) \end{array} \tag{91}$$

be a transition density (i.e. $P_{\rm out}$ is a measurable function such that for all $x \in \mathbb{R}$, $\int_{\mathbb{R}} P_{\rm out}(y|x)dy = 1$). Recall the free entropy expression (16) of the scalar channel (15). Let $\langle - \rangle$ denotes the expectation operator w.r.t the posterior distribution of $P(w = W^*|Y,V)$ and let w be drawn from this posterior.

***Lemma** B.1:* $\Psi_{P_{\rm out}}$ is a convex, non-decreasing function on $[0,\rho]$. For all $0 < q < \rho$,

$$\Psi'_{P_{\rm out}}(q) = \frac{1}{2(\rho-q)}\mathbb{E}\langle w W^*\rangle = \frac{1}{2(\rho-q)}\mathbb{E}[\langle w\rangle^2]\,, \tag{92}$$

$$\Psi''_{P_{\rm out}}(q) = \frac{1}{2(\rho-q)^2}\mathbb{E}\big[\big(\langle w^2\rangle - \langle w\rangle^2 - 1\big)^2\big]\,. \tag{93}$$

Proof: Let us define $X = \sqrt{q}\,V + \sqrt{\rho-q}\,W^*$. Then

$$\Psi_{P_{\rm out}}(q) = \mathbb{E}\int dX \frac{1}{\sqrt{2\pi(\rho-q)}} e^{-\frac{(X-\sqrt{q}\,V)^2}{2(\rho-q)}} \int dY P_{\rm out}(Y|X) \ln \int dx \frac{1}{\sqrt{2\pi(\rho-q)}} e^{-\frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)}} P_{\rm out}(Y|x)\,. \tag{94}$$

Using this expression, one can verify that $\Psi_{P_{\rm out}}$ is indeed continuous on $[0,\rho]$. It remains to show that the second derivative of $\Psi_{P_{\rm out}}$ is non-negative on $(0,\rho)$. Let us compute the derivatives of $\Psi_{P_{\rm out}}$ for $0 < q < \rho$:

$$\begin{aligned}
\Psi'_{P_{\rm out}}(q) &= \mathbb{E}\Big[\Big(\frac{1}{2(\rho-q)} - \frac{(X-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(X-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big) \ln \int dx \frac{1}{\sqrt{2\pi(\rho-q)}} e^{-\frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)}} P_{\rm out}(Y|x)\Big] \\
&\quad + \mathbb{E}\Big\langle \frac{1}{2(\rho-q)} - \frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(x-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big\rangle \\
&= \mathbb{E}\Big\langle \frac{(X-\sqrt{q}\,V)}{2(\rho-q)} \frac{(x-\sqrt{q}\,V)}{\rho-q}\Big\rangle + \mathbb{E}\Big[\frac{1}{2(\rho-q)} - \frac{(X-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(X-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big] \\
&= \frac{1}{2(\rho-q)}\mathbb{E}\langle w W^*\rangle \\
&= \frac{1}{2(\rho-q)}\mathbb{E}[\langle w\rangle^2]\,,
\end{aligned} \tag{95}$$

where we used the Nishimori property (Proposition A.1) and Gaussian integrations by parts w.r.t V. Let now compute the second derivative.

$$\begin{aligned}
2\Psi''_{P_{\rm out}}(q) &= \frac{\partial}{\partial q}\Big[\frac{1}{(\rho-q)^2}\mathbb{E}\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big] \\
&= \frac{2}{(\rho-q)^3}\mathbb{E}\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle - \frac{1}{(\rho-q)^2\sqrt{q}}\mathbb{E}[V(X-\sqrt{q}\,V)] \\
&\quad + \frac{2}{(\rho-q)^2}\mathbb{E}\Big[\Big(\frac{1}{2(\rho-q)} - \frac{(X-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(X-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big)\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big] \\
&\quad - \frac{1}{(\rho-q)^2}\mathbb{E}\Big[\Big\langle\frac{1}{2(\rho-q)} - \frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(x-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big]
\end{aligned}$$

$$\begin{aligned}
&= \frac{2}{(\rho-q)^2}\mathbb{E}\langle wW^*\rangle - \frac{1}{(\rho-q)^{3/2}\sqrt{q}}\mathbb{E}[VW^*] \\
&+ \frac{1}{(\rho-q)^4}\mathbb{E}\big[\big\langle (X-\sqrt{q}\,V)^2(x-\sqrt{q}\,V)^2\big\rangle\big] - \frac{1}{(\rho-q)^4}\mathbb{E}\big[\big\langle (X-\sqrt{q}\,V)^2(x-\sqrt{q}\,V)\big\rangle\big\langle (x-\sqrt{q}\,V)\big\rangle\big] \\
&- \frac{1}{(\rho-q)^3}\mathbb{E}\big[\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\big] - \frac{1}{(\rho-q)^3}\mathbb{E}[(X-\sqrt{q}\,V)^2] \\
&- \frac{1}{(\rho-q)^2}\mathbb{E}\Big[\Big\langle\frac{1}{2(\rho-q)} - \frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(x-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big] \\
&= \frac{1}{(\rho-q)^2}\mathbb{E}[\langle w\rangle]^2 + \frac{1}{(\rho-q)^2}\mathbb{E}[\langle w^2\rangle^2] - \frac{1}{(\rho-q)^2}\mathbb{E}[\langle w^2\rangle\langle w\rangle^2] - \frac{1}{(\rho-q)^2}\mathbb{E}[(W^*)^2] \\
&- \frac{1}{(\rho-q)^2}\mathbb{E}\Big[\Big\langle\frac{1}{2(\rho-q)} - \frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(x-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big]\,. \qquad (96)
\end{aligned}$$

Let us now compute the last term:

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle\frac{1}{2(\rho-q)} - \frac{(x-\sqrt{q}\,V)^2}{2(\rho-q)^2} + \frac{V(x-\sqrt{q}\,V)}{2\sqrt{q}(\rho-q)}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big] \\
&= -\mathbb{E}\Big[\Big\langle\frac{x-\sqrt{q}\,V}{2(\rho-q)}\Big\rangle\Big\langle\frac{x-\sqrt{q}\,V}{\rho-q}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\Big] - 2\mathbb{E}\Big[\Big\langle\frac{x-\sqrt{q}\,V}{2(\rho-q)}\Big\rangle\langle x-\sqrt{q}\,V\rangle\Big] \\
&+ 2\mathbb{E}\Big[\Big\langle\frac{x-\sqrt{q}\,V}{2(\rho-q)^2}\Big\rangle\big\langle (X-\sqrt{q}\,V)^2(x-\sqrt{q}\,V)\big\rangle\Big] - \mathbb{E}\Big[\Big\langle\frac{x-\sqrt{q}\,V}{2(\rho-q)^2}\Big\rangle\big\langle (X-\sqrt{q}\,V)(x-\sqrt{q}\,V)\big\rangle\langle x-\sqrt{q}\,V\rangle\Big] \\
&= -\mathbb{E}[\langle w\rangle^4] - \mathbb{E}[\langle w\rangle^2] + 2\mathbb{E}[\langle w^2\rangle\langle w\rangle^2]\,. \qquad (97)
\end{aligned}$$

Putting all together:

$$\begin{aligned}
\Psi''_{P_{\rm out}}(q) &= \frac{1}{2(\rho-q)^2}\big(\mathbb{E}[\langle w\rangle^2] + \mathbb{E}[\langle w^2\rangle^2] - \mathbb{E}[\langle w^2\rangle\langle w\rangle^2] - \mathbb{E}[(W^*)^2] + \mathbb{E}[\langle w\rangle^4] + \mathbb{E}[\langle w\rangle^2] - 2\mathbb{E}[\langle w^2\rangle\langle w\rangle^2]\big) \\
&= \frac{1}{2(\rho-q)^2}\big(\mathbb{E}[\langle w\rangle^2] + \mathbb{E}[(\langle w^2\rangle - \langle w\rangle^2)^2] - \mathbb{E}[\langle w^2\rangle - \langle w\rangle^2]\big) \\
&= \frac{1}{2(\rho-q)^2}\big(\mathbb{E}[(\langle w^2\rangle - \langle w\rangle^2)^2] - 2\mathbb{E}[\langle w^2\rangle - \langle w\rangle^2] + 1\big) \\
&= \frac{1}{2(\rho-q)^2}\mathbb{E}\big[\big(\langle w^2\rangle - \langle w\rangle^2 - 1\big)^2\big] \geq 0\,. \qquad (98)
\end{aligned}$$

■

Suppose now that $P_{\rm out}$ corresponds to the channel (57). Under hypothesis (H2) one can differentiate $\Psi_{P_{\rm out}}$ in order to obtain (using the Nishimori identity for the second equality):

Lemma B.2: Suppose that hypotheses (h1) and (H2) hold. Then for all $q \in [0,\rho]$,

$$\begin{aligned}
\Psi'_{P_{\rm out}}(q) &= \frac{1}{2}\mathbb{E}\big\langle u'_Y(\sqrt{q}\,V + \sqrt{\rho-q}\,w)u'_Y(\sqrt{q}\,V + \sqrt{\rho-q}\,W^*)\big\rangle \\
&= \frac{1}{2}\mathbb{E}\big[\big\langle u'_Y(\sqrt{q}\,Z + \sqrt{1-q}\,w)\big\rangle^2\big] \geq 0\,, \qquad (99)
\end{aligned}$$

where we used the notation $u_y(x) = \log P_{\rm out}(y|x)$. In particular, $\Psi'_{P_{\rm out}}$ is bounded.

Appendix C
A general class of models satisfying the hypothesis

Suppose that for all $(z,a) \in \mathbb{R} \times \mathbb{R}^{k_A}$, $|\varphi(z,a)| \leq c_1 + c_2|z|^p$ for some constants $p \geq 1$ and $c_1, c_2 \geq 0$. Then, by Jensen's inequality:

$$\mathbb{E}\Big[\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1, A_1\Big)^{2+\gamma}\Big] \leq 2^{1+\gamma}c_1^{2+\gamma} + 2^{1+\gamma}c_2^{2+\gamma}\,\mathbb{E}\Big[\Big|\frac{[\mathbf{\Phi X}^*]_1}{\sqrt{n}}\Big|^{p(2+\gamma)}\Big]\,. \tag{100}$$

Notice that $[\mathbf{\Phi X}^*]_1$ is equal in law to $\|\mathbf{X}^*\|Z$, where $Z \sim \mathcal{N}(0,1)$ is indepent of $\mathbf{X}^*$. Then, by Jensen's inequality:

$$\mathbb{E}\Big[\Big|\frac{[\mathbf{\Phi X}^*]_1}{\sqrt{n}}\Big|^{p(2+\gamma)}\Big] = \mathbb{E}\big[|Z^{p(2+\gamma)}|\big]\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n (X_i^*)^2\Big)^{p(1+\gamma/2)}\Big] \leq \mathbb{E}\big[|Z^{p(2+\gamma)}|\big]\mathbb{E}\big[(X_1^*)^{p(2+\gamma)}\big]\,. \tag{101}$$

Thus (h2) is satisfied as soon as $\mathbb{E}\left[(X_1^*)^{p(2+\gamma)}\right] < \infty$.

Appendix D
Proof of Proposition 5.1

We will first prove the following lemma:

Lemma *D.1:*

$$\begin{aligned} f_t' = &-\frac{1}{2}\mathbb{E}\left\langle \left(\frac{1}{n}\sum_{\mu=1}^m u_{Y_\mu}'(S_{t,\mu})u_{y_\mu}'(s_{t,\mu}) - r\right)\left(\frac{1}{n}\sum_{i=1}^n X_i^* x_i - q(t)\right)\right\rangle_t + \frac{1}{2}rq(t) - \frac{\rho r}{2} \\ &-\frac{1}{2}\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''(Y_\mu|S_{t,\mu})}{P(Y_\mu|S_{t,\mu})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i^*)^2 - \rho\right)\frac{1}{n}\log\left(\mathcal{Z}_t\right)\right] \end{aligned} \tag{102}$$

Proof: We start by differentiating the Hamiltonian:

$$\mathcal{H}_t'(\mathbf{y},\mathbf{y}',\mathbf{x},\mathbf{w}) = -\sum_{\mu=1}^n s_{t,\mu}' u_{y_\mu}'\left(s_{t,\mu}\right) - \frac{\sqrt{r}}{2\sqrt{t}}\sum_{i=1}^n x_i(y_i' - \sqrt{tr}x_i)$$

By definition

$$f_t = \frac{1}{n}\mathbb{E}_{\mathbf{\Phi}}\int d\mathbf{Y}d\mathbf{Y}'dP_0(\mathbf{X}^*)D\mathbf{W}^* e^{-\mathcal{H}_t(\mathbf{Y},\mathbf{Y}',\mathbf{X}^*,\mathbf{W}^*)}\log\left(\int dP_0(\mathbf{x})D\mathbf{w}\, e^{-\mathcal{H}_t(\mathbf{Y},\mathbf{Y}',\mathbf{x},\mathbf{w})}\right)$$

so that the derivative of the interpolating free entropy reads, for $0 < t < 1$,

$$f_t' = \underbrace{\frac{1}{n}\mathbb{E}\left[\mathcal{H}_t'(\mathbf{Y},\mathbf{Y}',\mathbf{X}^*,\mathbf{W}^*)\log\left(\mathcal{Z}_t\right)\right]}_{A} + \underbrace{\frac{1}{n}\mathbb{E}\Big\langle \mathcal{H}_t'(\mathbf{Y},\mathbf{Y}',\mathbf{x},\mathbf{w})\Big\rangle_t}_{B}$$

Let us compute A. Let $1 \leq \mu \leq m$.

$$\mathbb{E}\left[S_{t,\mu}' u_{Y_\mu}'(S_{t,\mu})\log\left(\mathcal{Z}_t\right)\right] = \mathbb{E}\left[\left(-\frac{[\mathbf{\Phi X}^*]_\mu}{2\sqrt{n(1-t)}} + \frac{q(t)}{2\sqrt{\int_0^s q(s)ds}}V_\mu + \frac{\rho - q(t)}{2\sqrt{\int_0^s (\rho - q(s))ds}}W_\mu^*\right)u_{Y_\mu}'(S_{t,\mu})\log\left(\mathcal{Z}_t\right)\right]$$

Compute

$$\mathbb{E}\left[\frac{[\mathbf{\Phi X}^*]_\mu}{2\sqrt{n(1-t)}}u_{Y_\mu}'\left(S_{t,\mu}\right)\log\left(\mathcal{Z}_t\right)\right] = \frac{1}{2\sqrt{n(1-t)}}\sum_{i=1}^n \mathbb{E}\left[X_i^*\Phi_{\mu,i}u_{Y_\mu}'\left(S_{t,\mu}\right)\log\left(\mathcal{Z}_t\right)\right]$$

By Gaussian integration by parts with respect to the $\Phi_{\mu,i}$ we obtain

$$\begin{aligned} \mathbb{E}\left[\frac{[\mathbf{\Phi X}^*]_\mu}{\sqrt{n(1-t)}}u_{Y_\mu}'\left(S_{t,\mu}\right)\log\left(\mathcal{Z}_t\right)\right] &= \frac{1}{n}\sum_{i=1}^n\left(\mathbb{E}\left[(X_i^*)^2\left(u_{Y_\mu}''\left(S_{t,\mu}\right) + u_{Y_\mu}'\left(S_{t,\mu}\right)^2\right)\log\left(\mathcal{Z}_t\right)\right] + \mathbb{E}\Big\langle X_i^* x_i u_{Y_\mu}'\left(S_{t,\mu}\right)u_{y_\mu}'\left(s_{t,\mu}\right)\Big\rangle_t\right) \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (X_i^*)^2\frac{P_{\text{out}}''(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\log\left(\mathcal{Z}_t\right)\right] + \mathbb{E}\Big\langle\frac{1}{n}\sum_{i=1}^n X_i^* x_i u_{Y_\mu}'\left(S_{t,\mu}\right)u_{y_\mu}'\left(s_{t,\mu}\right)\Big\rangle_t \end{aligned} \tag{103}$$

Because $u''_{Y_\mu}(x) + u'_{Y_\mu}(x)^2 = \frac{P''_{\text{out}}(Y_\mu|x)}{P_{\text{out}}(Y_\mu|x)}$. Using again Gaussian integration by part and the previous formula, we obtain

$$\mathbb{E}\left[\left(\frac{q(t)}{\sqrt{\int_0^t q(s)ds}}V_\mu + \frac{\rho - q(t)}{\sqrt{\int_0^t (\rho - q(s))ds}}W^*_\mu\right) u'_{Y_\mu}(S_{\mu,t}) \log(\mathcal{Z}_t)\right] = \mathbb{E}\left[\rho \frac{P''_{\text{out}}(Y_\mu|S_{\mu,t})}{P_{\text{out}}(Y_\mu|S_{\mu,t})} \log(\mathcal{Z}_t)\right] + \mathbb{E}\left\langle q(t) u'_{Y_\mu}(S_{\mu,t}) u'_{Y_\mu}(s_{\mu,t})\right\rangle_t \tag{104}$$

Putting equations (103) and (104) together, we have

$$\begin{aligned}
&\mathbb{E}\left[S'_{t,\mu} u'_{Y_\mu}(S_{t,\mu}) \log(\mathcal{Z}_t)\right] \\
&\quad = -\frac{1}{2}\mathbb{E}\left[\frac{P''_{\text{out}}(Y_\mu|S_{\mu,t})}{P_{\text{out}}(Y_\mu|S_{\mu,t})}\left(\frac{1}{n}\sum_{i=1}^n (X^*_i)^2 - \rho\right)\log(\mathcal{Z}_t)\right] - \frac{1}{2}\mathbb{E}\left\langle\left(\frac{1}{n}\sum_{i=1}^n X^*_i x_i - q(t)\right) u'_{Y_\mu}(S_{t,\mu})\, u'_{y_\mu}(s_{t,\mu})\right\rangle_t
\end{aligned}$$

It remain to compute, using again the Gaussian integration by parts,

$$\begin{aligned}
\mathbb{E}\left[\frac{\sqrt{r}}{2\sqrt{t}}\sum_{i=1}^n X^*_i (Y'_i - \sqrt{tr} X^*_i)\log(\mathcal{Z}_t)\right] &= \mathbb{E}\left[\frac{\sqrt{r}}{2\sqrt{t}}\sum_{i=1}^n X^*_i Z'_i \log(\mathcal{Z}_t)\right] \\
&= \mathbb{E}\left[\frac{r}{2}\sum_{i=1}^n X^*_i \Big\langle (x_i - X^*_i - Z'_i)\Big\rangle_t\right] \\
&= \frac{r}{2}\mathbb{E}\Big\langle \sum_{i=1}^n X_i x_i \Big\rangle_t - n\frac{\rho r}{2}
\end{aligned}$$

Therefore

$$\begin{aligned}
A = &-\frac{1}{2}\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{\mu,t})}{P_{\text{out}}(Y_\mu|S_{\mu,t})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)\frac{1}{n}\log(\mathcal{Z}_t)\right] + \frac{1}{2} r q(t) - \frac{r\rho}{2} \\
&- \frac{1}{2}\mathbb{E}\left\langle\left(\frac{1}{n}\sum_{i=1}^n X^*_i x_i - q(t)\right)\left(\frac{1}{n}\sum_{\mu=1}^m u'_{Y_\mu}(S_{t,\mu})\, u'_{y_\mu}(s_{t,\mu}) - r\right)\right\rangle_t
\end{aligned}$$

To obtain the Lemma, it remain to show that $B = 0$. This is a consequence of the Nishimori identity (see Appendix A):

$$B = \frac{1}{n}\mathbb{E}\Big\langle \mathcal{H}'_t(\mathbf{Y}, \mathbf{Y}', \mathbf{x}, \mathbf{w})\Big\rangle_t = \frac{1}{n}\mathbb{E}\Big[\mathcal{H}'_t(\mathbf{Y}, \mathbf{Y}', \mathbf{X}^*, \mathbf{W}^*)\Big] = 0$$

■

Lemma *D.2:* Under conditions (H1) and (H2)

$$\mathbb{E}\left[\frac{1}{\sqrt{m}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)\frac{1}{n}\log(\mathcal{Z}_t)\right] \xrightarrow[n\to\infty]{} 0 \tag{105}$$

uniformly in $t \in [0,1]$.

Proof: By the Cauchy-Schwarz inequality,

$$\begin{aligned}
&\left|\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)\frac{1}{n}\log(\mathcal{Z}_t)\right] - \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right) f_t\right]\right| \\
&\le \left(\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\right)^2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)^2\right]\mathbb{E}\left[\left(\frac{1}{n}\log(\mathcal{Z}_t) - f_t\right)^2\right]\right)^{1/2}
\end{aligned}$$

Conditionnaly to X, the Y_μ are independent, identically distributed and centered. Therefore

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\right)^2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{\mu=1}^m \frac{P''_{\text{out}}(Y_\mu|S_{t,\mu})}{P_{\text{out}}(Y_\mu|S_{t,\mu})}\right)^2\middle| X\right]\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)^2\right] \\
&= \frac{m}{n}\mathbb{E}\left[\mathbb{E}\left[\left(\frac{P''_{\text{out}}(Y_1|S_{t,1})}{P_{\text{out}}(Y_1|S_{t,1})}\right)^2\middle| X\right]\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (X^*_i)^2 - \rho\right)^2\right]
\end{aligned}$$

Under condition (H2), there exists a constant $C > 0$ such that

$$\mathbb{E}\left[\left(\frac{P''_{\text{out}}(Y_1|S_{t,1})}{P_{\text{out}}(Y_1|S_{t,1})}\right)^2\middle| X\right] \le C$$

Consequently, $\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum\limits_{\mu=1}^{m}\frac{P''_{\mathrm{out}}(Y_\mu|S_{t,\mu})}{P_{\mathrm{out}}(Y_\mu|S_{t,\mu})}\right)^2\left(\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}(X_i^*)^2-\rho\right)^2\right]$ is upper bounded by a constant. By Theorem H.1 we have $\mathbb{E}\big[(\frac{1}{n}\log(\mathcal{Z}_t)-f_t)^2\big]\xrightarrow[n\to\infty]{}0$, uniformly in $t\in[0,1]$. The lemma follows. ■

Appendix E
Boundedness of an overlap fluctuation

In this appendix we show that the "overlap fluctuation"

$$\mathbb{E}\Big\langle\Big(\frac{1}{n}\sum_{\mu=1}^{m}u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu})-r_n\Big)^2\Big\rangle_t\leq 2r_n^2+2\mathbb{E}\Big\langle\Big(\frac{1}{n}\sum_{\mu=1}^{m}u'_{Y_{t,\mu}}(S_{t,\mu})u'_{Y_{t,\mu}}(s_{t,\mu})\Big)^2\Big\rangle_t \tag{106}$$

is bounded uniformly in t under hypothesis (H2) on φ. From the representation (3)

$$\begin{aligned}u_{Y_{t,\mu}}(s)&=\ln P_{\mathrm{out}}(Y_{t,\mu}|s)\\&=\ln\int dP_A(a_\mu)(2\pi)^{-1/2}e^{-\frac{1}{2}(Y_{t,\mu}-\varphi(s,a_\mu))^2}\end{aligned}$$

so

$$u'_{Y_{t,\mu}}(s)=\frac{\int dP_A(a_\mu)(Y_{t,\mu}-\varphi(s,a_\mu))\varphi'(s,a_\mu)e^{-\frac{1}{2}(Y_{t,\mu}-\varphi(s,a_\mu))^2}}{\int dP_A(a_\mu)e^{-\frac{1}{2}(Y_{t,\mu}-\varphi(s,a_\mu))^2}}$$

where φ' is the derivative w.r.t the first argument. From (1) we get $|Y_{t,\mu}|\leq\sup|\varphi|+|Z_\mu|$ we immediately obtain for all $s\in\mathbb{R}$

$$|u'_{Y_{t,\mu}}(s)|\leq(2\sup|\varphi|+|Z_\mu|)\sup|\varphi'| \tag{107}$$

where the supremum is taken over both arguments of φ. From (107) and 106 we see that it suffices to check that

$$\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{\mu=1}^{m}(2\sup|\varphi|+Z_\mu)^2\Big)^2\Big]\leq C(\varphi)$$

where $C(\varphi)$ is a constant depending only on φ. This is easily seen by expanding all squares and using that m/n is bounded.

Appendix F
Approximation

In this section, we suppose that Theorem 2.1 holds for channels of the form (1) under the hypotheses (H1) and (H2).

We show in this section that this imply that Theorem 2.1 holds under the hypotheses (h1) and (h2). We start by relaxing the hypothesis (H1).

***Proposition** F.1:* Suppose that (h1) and (H2) hold. Then Theorem 2.1 holds.

Proof: The ideas are basically the same that in [26]. We omit the details here for the sake of brevity. ■

***Proposition** F.2:* Suppose that (h1) and (h2) hold. Then, Theorem 2.1 holds for the output channel (1).

To prove Proposition F.2 we will approximate the function φ with a function $\hat{\varphi}$ which is $\mathcal{C}^\infty$ with compact support. In the following, G is a standard Gaussian random variable, independent of everything else.

***Proposition** F.3:* Suppose that $\Big(\varphi(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1,A_1)\Big)_{n\geq1}$ is bounded in $L^{2+\gamma}$ for some $\gamma>0$. Then, for all $\epsilon>0$, there exist $\hat{\varphi}\in\mathcal{C}^\infty(\mathbb{R}\times\mathbb{R}^{k_A})$ with compact support, such that

$$\mathbb{E}\left[(\varphi(\sqrt{\rho}G,A)-\hat{\varphi}(\sqrt{\rho}G,A))^2\right]\leq\epsilon$$

and for n large enough, we have

$$\mathbb{E}\left[\left(\varphi\left(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1,A_1\right)-\hat{\varphi}\left(\frac{1}{\sqrt{n}}[\mathbf{\Phi X}^*]_1,A_1\right)\right)^2\right]\leq\epsilon$$

Proof: Notice that $[\mathbf{\Phi X}^*]_1=\|\mathbf{X}^*\|G$ in law. Thus, if $|X_1^*|$ is a constant random variable, then φ is in $L^2(\mathbb{R}\times\mathbb{R}^{k_A})$ with the measure induced by $(\sqrt{\rho}G,A_1)$. The result follows by density of the $\mathcal{C}^\infty$ functions with compact support in L^2.

We consider now the case where $|X_1^*|$ is not constant. We start with a useful lemma.

***Lemma** F.4:* Let $f:\mathbb{R}\to\mathbb{R}_+$ be a measurable function. Let $G\sim\mathcal{N}(0,1)$ and $0<a\leq b$. Then

$$\mathbb{E}[f(\sqrt{a}G)]\leq\sqrt{\frac{b}{a}}\mathbb{E}[f(\sqrt{b}G)].$$

In particular, if $f(\sqrt{b}G)$ is integrable then $f(\sqrt{a}G)$ is also integrable.

Proof: For $x \in \{a, b\}$, a change of variables gives: $\sqrt{x}\mathbb{E}[f(\sqrt{x}G)] = (2\pi)^{-1/2} \int e^{-g^2/(2x)} f(g)dg$ which is clearly non-decreasing in x. ■

$|X_1^*|$ is not constant and $\rho = \mathbb{E}[(X_1^*)^2]$, therefore there exists $\rho < r < r'$ such that $\mathbb{P}(r \leq (X_1^*)^2 \leq r') > 0$. Consequently, using Lemma F.4,

$$\begin{aligned}\mathbb{P}(r \leq (X_1^*)^2 \leq r')\mathbb{E}[\varphi(\sqrt{r}G, A_1)^2] &= \mathbb{E}[1_{r \leq (X_1^*)^2 \leq r'} \varphi(\sqrt{r}G, A_1)^2] \\ &\leq \mathbb{E}[1_{r \leq (X_1^*)^2 \leq r'} \frac{|X_1^*|}{\sqrt{r}} \varphi(|X_1^*|G, A_1)^2] \leq \sqrt{\frac{r'}{r}} \mathbb{E}[\varphi(|X_1^*|G, A_1)^2] < \infty\end{aligned}$$

Therefore $\mathbb{E}[\varphi(\sqrt{r}G, A_1)^2] < \infty$.

Let $\epsilon > 0$. We have just proved that $\varphi \in L^2(\mathbb{R} \times \mathbb{R}^{k_A})$ with the measure induced by $(\sqrt{r}G, A_1)$. There exists a $\mathcal{C}^\infty$ function with compact support $\hat{\varphi}$ such that $\mathbb{E}\left[(\varphi(\sqrt{r}G, A) - \hat{\varphi}(\sqrt{r}G, A))^2\right] \leq \epsilon$. Thus by Lemma F.4

$$\mathbb{E}\left[(\varphi(\sqrt{\rho}Z, A) - \hat{\varphi}(\sqrt{\rho}Z, A))^2\right] \leq \sqrt{\frac{r}{\rho}} \mathbb{E}\left[(\varphi(\sqrt{r}Z, A) - \hat{\varphi}(\sqrt{r}Z, A))^2\right] \leq \sqrt{\frac{r}{\rho}} \epsilon$$

It remains to bound $\mathbb{E}\left[\left(\varphi\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right) - \hat{\varphi}\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right)\right)^2\right]$. By the law of large numbers, $\frac{1}{n}\|\mathbf{X}\|^2 \xrightarrow[n\to\infty]{\mathbb{P}} \rho$. Thus $\mathbb{P}(\frac{1}{n}\|\mathbf{X}\|^2 \notin [\rho/2, r]) \xrightarrow[n\to\infty]{} 0$. We now apply Hölder's inequality:

$$\begin{aligned}&\mathbb{E}\left[1_{\|\mathbf{X}\|^2/n \notin [\rho/2,r]} \left(\varphi\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right) - \hat{\varphi}\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right)\right)^2\right] \\ &\qquad \leq \mathbb{P}(\|\mathbf{X}\|^2/n \notin [\rho/2, r])^{\frac{\gamma}{2+\gamma}} \mathbb{E}\left[\left(\varphi\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right) - \hat{\varphi}\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right)\right)^{2+\gamma}\right]^{\frac{2}{2+\gamma}} \\ &\qquad \leq C\epsilon^{\frac{\gamma}{2+\gamma}}\end{aligned}$$

for some constant $C > 0$ and for n large enough. It remain to bound

$$\begin{aligned}&\mathbb{E}\left[1_{\|\mathbf{X}\|^2/n \in [\rho/2,r]} \left(\varphi\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right) - \hat{\varphi}\left(\frac{1}{\sqrt{n}}\|\mathbf{X}\|G, A_1\right)\right)^2\right] \\ &\qquad\qquad \leq \mathbb{E}\left[1_{\|\mathbf{X}\|^2/n \in [\rho/2,r]} \sqrt{\frac{nr}{\|\mathbf{X}\|^2}} \left(\varphi\left(\sqrt{r}G, A_1\right) - \hat{\varphi}\left(\sqrt{r}G, A_1\right)\right)^2\right] \\ &\qquad\qquad \leq \sqrt{\frac{2r}{\rho}} \mathbb{E}\left[\left(\varphi\left(\sqrt{r}G, A_1\right) - \hat{\varphi}\left(\sqrt{r}G, A_1\right)\right)^2\right] \\ &\qquad\qquad \leq \sqrt{\frac{2r}{\rho}} \epsilon\end{aligned}$$

■

In the remaining of this section, we prove Proposition F.2. Let $\epsilon > 0$. Let φ and $\hat{\varphi}$ as in Proposition F.2. Let f_n be the free entropy associated to φ and $\hat{f}_n$ the free entropy corresponding to $\hat{\varphi}$.

Lemma F.5: There exists a constant $C > 0$ such that for n large enough

$$|f_n - \hat{f}_n| \leq C\sqrt{\epsilon}$$

Proof: Consider the observation channel given by

$$\begin{cases} Y_{t,\mu} &= \sqrt{t}\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu, A_\mu\Big) + Z_\mu \\ \hat{Y}_{t,\mu} &= \sqrt{1-t}\hat{\varphi}\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu, A_\mu\Big) + \hat{Z}_\mu \end{cases}$$

for $1 \leq \mu \leq m$. Let $f_n(t)$ denote the interpolating free energy:

$$f_n(t) = \frac{1}{n}\mathbb{E}\log\int_{\mathbf{x},\mathbf{a}} dP_A(\mathbf{a})dP_0(\mathbf{x}) \exp\left(-\frac{1}{2}\sum_{\mu=1}^{m}\left(Y_{t,\mu} - \sqrt{t}\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{x}]_\mu, a_\mu\Big)\right)^2 + \left(\hat{Y}_{t,\mu} - \sqrt{1-t}\hat{\varphi}\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi}\mathbf{x}]_\mu, a_\mu\Big)\right)^2\right)$$

In order to shorten the notations, we will now write $\varphi_\mu^{(\mathbf{x},\mathbf{a})}$ and $\hat{\varphi}_\mu^{(\mathbf{x},\mathbf{a})}$ for $\varphi\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu, a_\mu\Big)$ and $\hat{\varphi}\Big(\frac{1}{\sqrt{n}}[\mathbf{\Phi x}]_\mu, a_\mu\Big)$. We compute, for $0 < t < 1$:

$$
\begin{aligned}
f_n'(t) &= \frac{m}{2n}\mathbb{E}\left[(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})})^2 - (\varphi_1^{(\mathbf{X}^*,\mathbf{A})})^2\right] + \frac{1}{2n}\sum_{\mu=1}^{m}\mathbb{E}\left\langle \varphi_\mu^{(\mathbf{x},\mathbf{a})}\varphi_\mu^{(\mathbf{X}^*,\mathbf{A})} - \hat{\varphi}_\mu^{(\mathbf{x},\mathbf{a})}\hat{\varphi}_\mu^{(\mathbf{X}^*,\mathbf{A})}\right\rangle_t \\
&= \frac{m}{2n}\mathbb{E}\left[(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})})^2 - (\varphi_1^{(\mathbf{X}^*,\mathbf{A})})^2\right] + \frac{m}{2n}\mathbb{E}\left\langle \varphi_1^{(\mathbf{x},\mathbf{a})}(\varphi_1^{(\mathbf{X}^*,\mathbf{A})} - \hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})}) + (\varphi_1^{(\mathbf{x},\mathbf{a})} - \hat{\varphi}_1^{(\mathbf{x},\mathbf{a})})\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})}\right\rangle_t
\end{aligned}
$$

We start by controlling the first term:

$$
\begin{aligned}
\left|\mathbb{E}\left[(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})})^2 - (\varphi_1^{(\mathbf{X}^*,\mathbf{A})})^2\right]\right| &\leq \left(\mathbb{E}\left[\left(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})} + \varphi_1^{(\mathbf{X}^*,\mathbf{A})}\right)^2\right]\mathbb{E}\left[\left(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})} - \varphi_1^{(\mathbf{X}^*,\mathbf{A})}\right)^2\right]\right)^{1/2} \\
&\leq C_0\sqrt{\epsilon}
\end{aligned}
$$

by Proposition F.3, for some constant C_0 and n large enough. The two other terms can be bounded the same way:

$$
\begin{aligned}
\left|\mathbb{E}\left\langle \varphi_1^{(\mathbf{x},\mathbf{a})}(\varphi_1^{(\mathbf{X}^*,\mathbf{A})} - \hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})})\right\rangle_t\right| &\leq \left(\mathbb{E}\left[\left(\varphi_1^{(\mathbf{X}^*,\mathbf{A})}\right)^2\right]\mathbb{E}\left[\left(\hat{\varphi}_1^{(\mathbf{X}^*,\mathbf{A})} - \varphi_1^{(\mathbf{X}^*,\mathbf{A})}\right)^2\right]\right)^{1/2} \\
&\leq C_0\sqrt{\epsilon}
\end{aligned}
$$

by Proposition F.3, for n large enough. Consequently, there exists a constant $C > 0$ such that for n large enough and for all $0 < t < 1$, $|f_n'(t)| \leq C\sqrt{\epsilon}$. Notice that $t \mapsto f_n(t)$ is continuous over $[0,1]$, $f_n(0) = \hat{f}_n$ and $f_n(1) = f_n$, hence $|f_n - \hat{f}_n| \leq \int_0^1 |f_n'(t)|dt \leq C\sqrt{\epsilon}$. ■

Let P_{out} denote the transition kernel associated to φ and $\hat{P}_{\text{out}}$ the one associated to $\hat{\varphi}$. Analogously to the previous Lemma, one can show:

Lemma F.6: There exists a constant $C' > 0$ such that for all $q \in [0,\rho]$

$$
|\Psi_{P_{\text{out}}}(q) - \Psi_{\hat{P}_{\text{out}}}(q)| \leq C'\sqrt{\epsilon}
$$

From there we obtain that

$$
\left|\sup_{r\geq 0}\inf_{q\in[0,\rho]} f_{RS}(q,r) - \sup_{r\geq 0}\inf_{q\in[0,\rho]} \hat{f}_{RS}(q,r)\right| \leq C'\sqrt{\epsilon} \quad \text{and} \quad \left|\sup_{q\in[0,\rho]}\inf_{r\geq 0} f_{RS}(q,r) - \sup_{q\in[0,\rho]}\inf_{r\geq 0} \hat{f}_{RS}(q,r)\right| \leq C'\sqrt{\epsilon} \tag{108}
$$

Applying Theorem 2.1 for P_{out}, we obtain that for n large enough $|f_n - \sup_{r\geq 0}\inf_{q\in[0,\rho]} f_{RS}(q,r)| \leq \epsilon$. We now combine this with (108) and Lemma F.5 we obtain that for n large enough

$$
\left|\hat{f}_n - \sup_{q\in[0,\rho]}\inf_{r\geq 0} \hat{f}_{RS}(q,r)\right| = \left|\hat{f}_n - \sup_{r\geq 0}\inf_{q\in[0,\rho]} \hat{f}_{RS}(q,r)\right| \leq (C + C')\sqrt{\epsilon} + \epsilon
$$

which concludes the proof of Proposition F.2.

Appendix G
A sup-inf formula

We first need to indroduce some notations about convex functions. Let f be a convex function on some interval $I \subset \mathbb{R}$. For $t \in I$ we will denote respectlively by $f'(t^-)$ and $f'(t^+)$ the left and right hand derivatives of f at t. We also define the subgradient of f at t as $\partial f(t) = [f'(t^-), f'(t^+)]$. If t is the right (resp. left) border of I, then we define $\partial f(t) = \{f'(t^-)\}$ (resp. $\partial f(t) = \{f'(t^+)\}$).

Lemma G.1: Let f and g be two convex, Lipschitz, non-decreasing functions on $\mathbb{R}_+$. For $q_1, q_2 \in \mathbb{R}_+$ we define $\psi(q_1,q_2) = f(q_1) + g(q_2) - q_1 q_2$. Then

$$
\sup_{q_1\geq 0}\inf_{q_2\geq 0}\psi(q_1,q_2) = \sup_{\substack{q_1\in\partial g(q_2)\\ q_2 = f'(q_1^+)}}\psi(q_1,q_2) = \sup_{\substack{q_1\in\partial g(q_2)\\ q_2\in\partial f(q_1)}}\psi(q_1,q_2)
$$

and these extremas are achieved at some (possibly many) couples. All these optimal couples are in $[0, \sup_{x\geq 0} g'(x^+)] \times [0, \sup_{x\geq 0} f'(x^+)]$. Moreover, if g is strictly convex, then the above extremas are achieved precisely on the same couples (q_1, q_2) and f is differentiable at q_1.

Corollary G.2: Let f be a convex, Lipschitz, non-decreasing function on $\mathbb{R}_+$. Define $\rho = \sup_{x\geq 0} f'(x^+)$. Let $g : [0,\rho] \to \mathbb{R}$ be a convex, Lipschitz, non-decreasing function. For $q_1 \in \mathbb{R}_+$ and $q_2 \in [0,\rho]$ we define $\psi(q_1,q_2) = f(q_1) + g(q_2) - q_1 q_2$. Then

$$
\sup_{q_1\geq 0}\inf_{q_2\in[0,\rho]}\psi(q_1,q_2) = \sup_{q_2\in[0,\rho]}\inf_{q_1\geq 0}\psi(q_1,q_2)
$$

Proof: In order to apply Lemma G.1 we need to extend g on $\mathbb{R}_+$. We thus define for $x \geq 0$

$$\tilde{g}(x) = \begin{cases} g(x) & \text{if } x \leq \rho \\ g(\rho) + (x-\rho) g'(\rho^+) & \text{if } x \geq \rho \end{cases}$$

$\tilde{g}$ is simply equal to g that we extend for $x \geq \rho$ using his tangent at ρ. Obviously $\tilde{g}$ is a convex, Lipschitz, non-decreasing function on $\mathbb{R}_+$. One can thus apply Lemma G.1:

$$\sup_{q_1 \geq 0} \inf_{q_2 \geq 0} f(q_1) + \tilde{g}(q_2) - q_1 q_2 = \sup_{q_2 \geq 0} \inf_{q_1 \geq 0} f(q_1) + \tilde{g}(q_2) - q_1 q_2$$

and both "sup-inf" are achieved on $[0, \sup_{x \geq 0} \tilde{g}'(x^+)] \times [0, \sup_{x \geq 0} f'(x^+)]$. By definition, $\rho = \sup_{x \geq 0} f'(x^+)$, thus

$$\begin{aligned} \sup_{q_1 \geq 0} \inf_{q_2 \in [0,\rho]} f(q_1) + \tilde{g}(q_2) - q_1 q_2 &= \sup_{q_1 \geq 0} \inf_{q_2 \geq 0} f(q_1) + \tilde{g}(q_2) - q_1 q_2 \\ &= \sup_{q_2 \geq 0} \inf_{q_1 \geq 0} f(q_1) + \tilde{g}(q_2) - q_1 q_2 = \sup_{q_2 \in [0,\rho]} \inf_{q_1 \geq 0} f(q_1) + \tilde{g}(q_2) - q_1 q_2 \end{aligned}$$

which concludes the proof because $\tilde{g}(q_2) = g(q_2)$ for $q_2 \in [0, \rho]$. ■

To prove Lemma G.1 we will need the following lemma on the Fenchel-Legendre transform.

Lemma G.3: Let $V \subset \mathbb{R}$ be an interval and let $g : V \to \mathbb{R}$ be a convex function. Define

$$g^* : x \in \mathbb{R} \mapsto \sup_{y \in V} \big\{ xy - g(y) \big\} \in \mathbb{R} \cup \{+\infty\}\,. \tag{109}$$

Let $D_{g^*} = \{x \in \mathbb{R} \,|\, g^*(x) < \infty\}$. Then g^* is a convex function on the interval $D_{g^*} \neq \emptyset$. Moreover, $D_{g^*} = \{a\}$ if and only if $g : x \mapsto ax$. For $x \in D_{g^*}$ the set of maximizers of (109) is of the form $[a_x, b_x]$, where $a_x, b_x \in \mathbb{R} \cup \{\pm\infty\}$. Then, the left-hand and right-hand derivatives of g^* at x are respectlively a_x and b_x.

In particular, if g is strictly convex then g^* is differentiable around every point in the interior of D_{g^*}.

Proof: g^* is convex because an supremum of linear functions is a convex function.

For $x \in \mathbb{R}$ we define the function $\varphi_x : y \mapsto xy - g(y)$. Let $x \in D_{g^*}$. We are only going to show that the left-hand derivative of g^* at x is equal to a_x, the result for the right-hand derivative is proved analogously.

G_x is then a closed interval of the form $[a, b]$ (b may be $+\infty$). Because of convexity, g^* is left- and right-hand differentiable at x. First of all, notice that if $a = -\infty$ then $g^*(x') = +\infty$ for any $x' < x$, so the left-hand derivative of g^* is not defined. We will thus concentrate on the cases where $a_x \in \mathbb{R}$ and $a = +\infty$.

Let us consider first the case where a_x is finite. By definition, $g^*(x) = \varphi_x(a)$. Let now $x' < x$. We have

$$a_x(x - x') = \varphi_x(a_x) - \varphi_{x'}(a_x) \geq g^*(x) - g^*(x')$$

which implies that $(g^*)'(x^-) \leq a_x$. The fact that a_x achieves the maximum in (109) implies that $x \in \partial g(a_x)$. Let now $x' < x$ in G_{g^*}. Notice that $b_{x'} \leq a_x$. We have then, by convexity $g(b_{x'}) \geq g(a_x) + (b_{x'} - a_x)x$ which implies

$$\begin{aligned} g^*(x) - g^*(x') &= a_x x - g(a_x) - b_{x'} x' + g(b_{x'}) \\ &\geq ax - g(a) - b_{x'} x' + g(a) + (b_{x'} - a_x)x \\ &\geq b_x (x - x') \end{aligned}$$

When $x' \to x$, one can verify easily that $b_{x'} \to a_x$. We obtain therefore that $(g^*)'(x^-) \leq a_x$. We conclude $(g^*)'(x^-) = a_x$.

Suppose now that $a_x = +\infty$. Let $x' < x'' < x$. $b_{x'}$ and $a_{x''}$ are necessarily finite, otherwise $g^*(x) = +\infty$. Applying the result we just proved, we have

$$\frac{g^*(x'') - g^*(x')}{x'' - x'} \xrightarrow[x' \to x'']{} a_{x''}$$

Thus, by convexity $(g^*)'(x^-) \geq a_{x''}$. Since $a_x = +\infty$, $a_{x''} \to +\infty$ when $x'' \to x$. We conclude $(g^*)'(x^-) = +\infty$. ■

Proof of Lemma G.1: Let $q_1, q_2 \geq 0$ such that $q_1 \in \partial g(q_2)$. Then, by convexity of g, $\psi(q_1, q_2) = \inf_{q_2' \geq 0} \psi(q_1, q_2')$. Consequently:

$$\sup_{q_1 \geq 0} \inf_{q_2 \geq 0} \psi(q_1, q_2) \geq \sup_{\substack{q_1 \in \partial g(q_2) \\ q_2 \in \partial f(q_1)}} \psi(q_1, q_2) \geq \sup_{\substack{q_1 \in \partial g(q_2) \\ q_2 = f'(q_1^+)}} \psi(q_1, q_2)$$

Let us now prove the converse bound. Let $L_g = \sup_{y \geq 0} g'(y^+)$. L_g is finite because g is Lipschitz. Notice that for $0 \leq y < L_g$, $g^*(y) < \infty$ while for $y > L_g$, $g^*(y) = +\infty$. f is continuous on $[0, L_g]$ and g^* is convex on $[0, L_g]$ ($g^*(L_g)$ may be equal to $+\infty$). Therefore $f - g^*$ achieves its suppremum at some $q_1^* \in [0, L_g]$.

If $0 < q_1^* < L_g$ then the optimality condition at q_1^* gives $f'(q_1^{*-}) - g^{*\prime}(q_1^{*-}) \geq 0$ and $f'(q_1^{*+}) - g^{*\prime}(q_1^{*+}) \leq 0$. Thus

$$g^{*\prime}(q_1^{*-}) \leq f'(q_1^{*-}) \leq f'(q_1^{*+}) \leq g^{*\prime}(q_1^{*+}) \tag{110}$$

We know by Lemma G.3 that $[g^{*\prime}(q_1^{*-}), g^{*\prime}(q_1^{*+})]$ is the set of maximizers of $q_2 \mapsto q_1^* q_2 - g(q_2)$. Consequently $q_2 = f'(q_1^{*+})$ maximizes $q_2 \mapsto q_1^* q_2 - g(q_2)$. This gives $q_1^* \in \partial g(q_2^*)$. We conclude:

$$\sup_{q_1 \geq 0} \inf_{q_2 \geq 0} \psi(q_1, q_2) = f(q_1^*) + g(q_2^*) - q_1^* q_2^* \leq \sup_{\substack{q_1 \in \partial g(q_2) \\ q_2 = f'(q_1^+)}} f(q_1) + g(q_2) - q_1 q_2$$

If now $q_1^* = 0$, the optimality condition is now: $f'(0^+) - g^{*\prime}(0^+) \leq 0$. f is non-decreasing, hence $0 \leq f'(0^+) \leq g^{*\prime}(0^+)$. g is non-decreasing, thus the set of maxizers of $g^*(0)$ is, by Lemma G.3 $[0, g^{*\prime}(0^+)]$. Therefore $q_2^* = f(0^+)$ maximizes $q_2 \mapsto q_1^* q_2 - g(q_2)$ and we conclude similarly as before.
If $q_1^* = L_g$, then the optimality condition is $f'(q_1^{*-}) - g^{*\prime}(q_1^{*-}) \geq 0$. Therefore $f'(q_1^+) \geq f'(q_1^-) \geq g^{*\prime}(q_1^{*-})$. Again by Lemma G.3, $q_2^* = f(q_1^{*+})$ maximizes $q_2 \mapsto q_1^* q_2 - g(q_2)$, we conclude similarly as before.

Suppose now that g is strictly convex. This means (by Lemma G.3) that g^* is differentiable. Let $(q_1, q_2) \in \mathbb{R}_+^2$ be a couple that achieves the supremum in one of the last two sup. This means in particular that $q_1 \in \partial g(q_2)$. By convexity of g we have then $\psi(q_1, q_2) = \inf_{q_2' \geq 0} \psi(q_1, q_2')$. Since $\psi(q_1, q_2) = \sup_{q_1' \geq 0} \inf_{q_2' \geq 0} \psi(q_1', q_2')$, (q_1, q_2) achieves the sup-inf. Let (q_1, q_2) be a couple that achieves the sup-inf. It remains to show that $q_1 \in \partial g(q_2)$ and $q_2 = f_1'(q_1^+)$. First of all, the fact that $\psi(q_1, q_2) = \inf_{q_2' \geq 0} \psi(q_1, q_2')$ implies (by convexity) that $q_1 \in \partial g(q_2)$. Now, as in the proof above we have to distiguish whenever $q_1 \in (0, L_g)$, $q_1 = 0$ or $q_1 = L_g$. When $q_1 \in (0, L_g)$, the strict convexity of g implies the differentiability of g^*, therefore the inequality (110) gives that f is differentiable at q_1 and $f'(q_1) = g^{*\prime}(q_1) = q_2$. The case $q_1 = 0$ goes the same way. It remains to see that q_1 could not be equal to L_g. Indeed, by strict convexity, the infimum of $q_2' \mapsto \psi(L_g, q_2')$ is only achieved when $q_2' \to +\infty$. Lemma G.3 gives then that $g^*(L_g^-) = +\infty$, which gives that the function $q_1 \mapsto f(q_1) - g^*(q_1)$ is decreasing on $[L_g - \epsilon, L_g]$, for some $\epsilon > 0$. The supremum can therefore not be achieved at $q_1 = L_g$. ■

The proof of Lemma G.1 could be straigthforwardly adapted, to obtain the next lemma.

Lemma G.4: Let g be a strictly convex, differentiable, Lipschitz non-decreasing function on $\mathbb{R}_+$. Define $\rho = \sup_{x \geq 0} g'(x)$. Let f be a convex, non-decreasing function on $[0, \rho]$, differentiable on $[0, \rho)$. For $q_1, q_2 \in \mathbb{R}_+$ we define $\psi(q_1, q_2) = f(q_1) + g(q_2) - q_1 q_2$. Then

$$\sup_{q_1 \in [0,\rho]} \inf_{q_2 \geq 0} \psi(q_1, q_2) = \sup_{\substack{q_1 = g'(q_2) \\ q_2 = f'(q_1)}} \psi(q_1, q_2)$$

Moreover, the above extremas are achieved precisely on the same couples.

Appendix H
Concentration of the free entropy

The goal of this Appendix is to prove the following concentration result. To simplify the notations we use $C(\varphi, S, \alpha)$, for a generic strictly positive constant depending *only* on φ, S and α (S the supremum over signal values). It is also understood that n and m are large enough and $m/n \to \alpha$

Theorem H.1: Suppose we have a prior with bounded support and $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is bounded and differentiable with respect to its firs argument with bounded derivative. We can find a constant $C(\varphi, S, \alpha) > 0$ such that

$$\mathbb{P}\big(|\frac{1}{n} \ln \mathcal{Z}_t - \mathbb{E}[\frac{1}{n} \ln \mathcal{Z}_t]| > r\big) \leq e^{-C(\varphi, S, \alpha) r^2 n} \tag{111}$$

for any $r > 0$. Moreover

$$\mathbb{E}\big[|\frac{1}{n} \ln \mathcal{Z}_t - \mathbb{E}[\frac{1}{n} \ln \mathcal{Z}_t]|^2\big] \leq \frac{1}{nC(\varphi, S, \alpha)} \tag{112}$$

We first recall some set-up and notation for the convenience of the reader. The interpolating Hamiltonian (62)-(60) is (we indicate only the annealed variables in its arguments)

$$-\sum_{\mu=1}^{m} \ln P_{\text{out}}\left(Y_\mu | s_{t,\mu}(\mathbf{x}, w_\mu)\right) + \frac{1}{2} \sum_{i=1}^{n} (Y_i' - \sqrt{tr} x_i)^2 \tag{113}$$

where

$$s_{t,\mu}(\mathbf{x}, w_\mu) = \sqrt{\frac{1-t}{n}} [\mathbf{\Phi x}]_\mu + k_1(t) V_\mu + k_2(t) w_\mu, \quad k_1(t) = \sqrt{\int_0^t q(v) dv}, \quad k_2(t) = \sqrt{\int_0^t (\rho - q(v)) dv}$$

We find it convenient to use the random function representation for the interpolating model, namely

$$\begin{cases} Y_{t,\mu} = \varphi\Big(\sqrt{\frac{1-t}{n}} [\mathbf{\Phi X}^*]_\mu + k_1(t) V_\mu + k_2(t) W_\mu^*, A_\mu\Big) + Z_\mu, \\ Y_{t,i}' = \sqrt{rt} X_i^* + Z_i' \end{cases}$$

where $\varphi(x,y)$ is bounded, in $\mathcal{C}^2$ with respect to x, and $\sup_{x,y}|\partial_x\varphi(x,y)|<\infty$, $\sup_{x,y}|\partial_x^2\varphi(x,y)|<\infty$. We will use the notation $\varphi'(x,y)=\partial_x\varphi(x,y)$, $\varphi''(x,y)=\partial_x^2\varphi(x,y)$. In this representation the iid random variables $A_\mu\sim P_A$ are arbitrary, and $Z_\mu\sim\mathcal{N}(0,1)$, $\mu=1,\dots,M$. We have (here $a_\mu\sim P_A$)

$$P_{t,\text{out}}(Y_{t,\mu}|s_{t,\mu}(\mathbf{x},\mathbf{w}))=\int dP_A(a_\mu)\,\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}\big(Y_{t,\mu}-\varphi(s_{t,\mu}(\mathbf{x},w_\mu),a_\mu\big)^2\}$$
$$=\int dP_A(a_\mu)\,\frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{1}{2}\Gamma_{t,\mu}(\mathbf{x},w_\mu,a_\mu)^2\Big\} \tag{114}$$

where, using the random function representation,

$$\Gamma_{t,\mu}(\mathbf{x},w_\mu,a_\mu)=\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu+k_1(t)V_\mu+k_2(t)W_\mu^*,A_\mu\Big)-\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{x}]_\mu+k_1(t)V_\mu+k_2(t)w_\mu,a_\mu\Big)+Z_\mu\,. \tag{115}$$

From (113), (114), (115) we can express the free entropy of the interpolating model as

$$\frac{1}{n}\ln\mathcal{Z}_t=\frac{1}{n}\ln\Big\{\int dP_0(\mathbf{x})dP_A(\mathbf{a})D\mathbf{w}\,e^{-\mathcal{H}_t(\mathbf{x},\mathbf{w},\mathbf{a})}\Big\} \tag{116}$$

where we introduced an "effective" Hamiltonian is now (we drop an irrelevant constant $M\sqrt{2\pi}$)

$$\mathcal{H}_t(\mathbf{x},\mathbf{w},\mathbf{a})=\frac{1}{2}\sum_{\mu=1}^{m}\Gamma_{t,\mu}(\mathbf{x},w_\mu,\mathbf{a}_\mu)^2+\frac{1}{2}\sum_{i=1}^{n}(\sqrt{rt}X_i^*+Z_i'-\sqrt{tr}x_i)^2\,. \tag{117}$$

The interpretation here is that $\mathbf{x},\mathbf{w},\mathbf{a}$ are *annealed* variables and $\mathbf{\Phi},\mathbf{V},\mathbf{A},\mathbf{Z},\mathbf{Z}',\mathbf{X}^*,\mathbf{W}^*$ are *quenched*. The inference problem is to recover $\mathbf{X}^*,\mathbf{W}^*$ given all other quenched variables.

Our goal is to show that the free energy (116) concentrates with respect to *all* quenched variables. We will first show concentration w.r.t all Gaussian variables $\mathbf{\Phi},\mathbf{V},\mathbf{Z},\mathbf{Z}',\mathbf{W}^*$ thanks to the classical Gaussian concentration inequality, then the concentration w.r.t $\mathbf{A}$ and finally the one w.r.t $\mathbf{X}^*$ thanks to Mc-Diarmid's inequality. The order in which we prove the concentrations matters. Here is a statement of these two inequalities.

***Proposition** H.2 (Tsirelson - Ibragimov - Sudakov inequality):* Let $\mathbf{U}=(U_1,\dots,U_N)$ be a vector of N independent standard normal random variables. Let $L>0$ and let $g:\mathbb{R}^N\to\mathbb{R}$ be a L-Lipschitz function with respect to the Euclidean distance. Then for any $r>0$,

$$\mathbb{P}\big(g(\mathbf{U})-\mathbb{E}g(\mathbf{U})\geq r\big)\leq e^{-\frac{r^2}{2L^2}}\,. \tag{118}$$

***Remark** H.3:* If g is differentiable and $\sup_{\mathbb{R}^N}\|\nabla g\|\leq L<\infty$ then g is Lipschitz w.r.t the Euclidean distance.

***Proposition** H.4 (McDiarmid inequality):* Let $\mathcal{U}\subset\mathbb{R}$. Let $g:\mathcal{U}^N\to\mathbb{R}$ a function that satisfies the bounded difference property, i.e., there exists some constants $c_1,\dots,c_N\geq 0$ such that

$$\sup_{\substack{u_1,\dots u_N\in\mathcal{U}^N\\ u_i'\in\mathcal{U}}}|g(u_1,\dots,u_i,\dots,u_N)-g(u_1,\dots,u_i',\dots,u_N)|\leq c_i\,,\ \text{ for all } 1\leq i\leq N\,.$$

Let $\mathbf{U}=(U_1,\dots,U_N)$ be a vector of N independent random variables that takes values in $\mathcal{U}$. Then for all $r\geq 0$,

$$\mathbb{P}\big(g(\mathbf{U})-\mathbb{E}g(\mathbf{U})\geq r\big)\leq e^{-\frac{2r^2}{\sum_{i=1}^N c_i^2}}\,. \tag{119}$$

Before we proceed it is useful to remark

$$\frac{1}{n}\ln\mathcal{Z}_t=\frac{1}{n}\ln\hat{\mathcal{Z}}_t-\frac{1}{2n}\sum_{\mu=1}^{m}Z_\mu^2-\frac{1}{2n}\sum_{i=1}^{n}Z_i'^2 \tag{120}$$

where

$$\frac{1}{n}\ln\hat{\mathcal{Z}}_t=\frac{1}{n}\ln\Big\{\int dP_0(\mathbf{x})dP_A(\mathbf{a})D\mathbf{w}\,e^{-\hat{\mathcal{H}}_t(\mathbf{x},\mathbf{w},\mathbf{a})}\Big\} \tag{121}$$

with

$$\hat{\mathcal{H}}_t(\mathbf{x},\mathbf{w},\mathbf{a})=\frac{1}{2}\sum_{\mu=1}^{m}\Big\{\hat{\Gamma}_{t,\mu}(\mathbf{x},w_\mu,\mathbf{a}_\mu)^2+2Z_\mu\hat{\Gamma}_{t,\mu}(\mathbf{x},w_\mu,\mathbf{a}_\mu)\Big\}+\frac{1}{2}\sum_{i=1}^{n}\Big\{(\sqrt{rt}X_i^*-\sqrt{tr}x_i)^2+2Z_i'(\sqrt{rt}X_i^*-\sqrt{tr}x_i)\Big\} \tag{122}$$

and

$$\hat{\Gamma}_{t,\mu}(\mathbf{x},w_\mu,a_\mu)=\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu+k_1(t)V_\mu+k_2(t)W_\mu^*,A_\mu\Big)-\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{x}]_\mu+k_1(t)V_\mu+k_2(t)w_\mu,a_\mu\Big).$$

Obviously, if (121) concentrates then (120) also concentrates. In the rest of the analysis we show concentration of (121).

A. Concentration with respect to Gaussian random variables Z_μ, Z'_i, V_μ, W^*_μ, $\Phi_{\mu i}$

We set $g = \frac{1}{n}\ln\hat{\mathcal{Z}}_t$ and first prove concentration with respect to Z_μ, Z'_i. We have for the gradient with respect to Z_μ and Z'_i

$$\|\nabla g\|^2 = \sum_{\mu=1}^{m}\big|\frac{\partial g}{\partial Z_\mu}\big|^2 + \sum_{i=1}^{n}\big|\frac{\partial g}{\partial Z'_i}\big|^2. \tag{123}$$

Each of these derivatives are of the form $\partial g = n^{-1}\langle\partial\hat{\mathcal{H}}_t\rangle_{\hat{\mathcal{H}}_t}$ where the Gibbs bracket $\langle -\rangle_{\hat{\mathcal{H}}_t}$ pertains to the effective Hamiltonian (122). We find

$$\begin{aligned}\big|\frac{\partial g}{\partial Z_\mu}\big| &= n^{-1}\big|\langle\hat{\Gamma}_{t,\mu}\rangle_{\hat{\mathcal{H}}_t}\big| \leq 2n^{-1}\sup|\partial_x\varphi| \\ \big|\frac{\partial g}{\partial Z'_i}\big| &= n^{-1}\big|\langle\sqrt{rt}X^*_i - \sqrt{tr}x_i\rangle_{\hat{\mathcal{H}}_t}\big| \leq 2n^{-1}S\end{aligned}$$

and replacing in (123)

$$\|\nabla g\|^2 \leq 2n^{-1}(\frac{m}{n}\sup|\partial_x\varphi| + S) \equiv L_n^2.$$

Applying Proposition (H.2) we have

$$\mathbb{P}\Big(\big|\frac{1}{n}\ln\hat{\mathcal{Z}}_t - \frac{1}{n}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}[\ln\hat{\mathcal{Z}}_t]\big| > r\Big) \leq 2e^{-C(\varphi,S,\alpha)nr^2} \tag{124}$$

where $\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}$ is the expectation w.r.t $\mathbf{Z},\mathbf{Z}'$ only, and $\mathbb{P}$ is the probability w.r.t all random variables.

Now we set $g = n^{-1}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}[\ln\hat{\mathcal{Z}}_t]$ and show concentration w.r.t the rest of the Gaussian variables, namely V_μ, W^*_μ, $\Phi_{\mu i}$. We have

$$\begin{aligned}\big|\frac{\partial g}{\partial V_\mu}\big| &= n^{-1}\big|\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[\Big\langle(\hat{\Gamma}_{t,\mu} + Z_\mu)\frac{\partial\hat{\Gamma}_{t,\mu}}{\partial V_\mu}\Big\rangle_{\mathcal{H}_t}\Big]\big| \\ &\leq n^{-1}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[(2\sup|\varphi| + |Z_\mu|)(2\sup|\partial_x\varphi|)\Big] \\ &= n^{-1}(2\sup|\varphi| + \sqrt{\frac{2}{\pi}})(2\sup|\partial_x\varphi|)\end{aligned}$$

The same inequality holds for $\big|\frac{\partial g}{\partial W^*_\mu}\big|$. To compute the derivative w.r.t $\Phi_{\mu i}$ we first remark

$$\begin{aligned}\frac{\partial\hat{\Gamma}_{t,\mu}}{\partial\Phi_{\mu i}} =& \sqrt{\frac{1-t}{n}}(X^*_i - x_i)\Big\{\partial_x\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu + k_1(t)V_\mu + k_2(t)W^*_\mu, A_\mu\Big) \\ &- \partial_x\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big)\Big\}.\end{aligned}$$

Therefore

$$\begin{aligned}\big|\frac{\partial g}{\partial\Phi_{\mu i}}\big| &= n^{-1}\big|\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[\Big\langle(\hat{\Gamma}_{t,\mu} + Z_\mu)\frac{\partial\hat{\Gamma}_{t,\mu}}{\partial\Phi_{\mu i}}\Big\rangle_{\mathcal{H}_t}\Big]\big| \\ &\leq n^{-3/2}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\Big[(2\sup|\varphi| + |Z_\mu|)(4S\sup|\partial_x\varphi|)\Big] \\ &= n^{-3/2}(2\sup|\varphi| + \sqrt{\frac{2}{\pi}})(4S\sup|\partial_x\varphi|)\end{aligned}$$

Putting these inequalities together we find for the gradient w.r.t V_μ, W_μ, $\Phi_{\mu i}$

$$\begin{aligned}\|\nabla g\|^2 &= \sum_{\mu=1}^{m}\big|\frac{\partial g}{\partial V_\mu}\big|^2 + \sum_{\mu=1}^{m}\big|\frac{\partial g}{\partial V_\mu}\big|^2 + \sum_{\mu=1}^{m}\sum_{i=1}^{n}\big|\frac{\partial g}{\partial\Phi_{\mu i}}\big|^2 \\ &\leq \frac{m}{n^2}(2\sup|\varphi| + \sqrt{\frac{2}{\pi}})(2\sup|\partial_x\varphi|) + \frac{mn}{n^3}(2\sup|\varphi| + \sqrt{\frac{2}{\pi}})(4S\sup|\partial_x\varphi|)\end{aligned}$$

and a direct application of (H.2) then yields

$$\mathbb{P}\Big(\big|\frac{1}{n}\mathbb{E}_{\mathbf{Z},\mathbf{Z}'}[\ln\hat{\mathcal{Z}}_t] - \frac{1}{n}\mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi}}[\ln\hat{\mathcal{Z}}_t]\big| > r\Big) \leq 2e^{-C(\varphi,S,\alpha)nr^2}. \tag{125}$$

By a simple application of the triangle inequality and the union bound estimates (124) and (125) imply

$$\mathbb{P}\Big(\big|\frac{1}{n}\ln\hat{\mathcal{Z}}_t - \frac{1}{n}\mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi}}[\ln\hat{\mathcal{Z}}_t]\big| > r\Big) \leq 4e^{-C(\varphi,S,\alpha)nr^2}. \tag{126}$$

where $\mathbb{P}$ is the probability w.r.t all random variables.

B. Bounded difference with respect to A_μ

The next step is an application of MacDiarmid's inequality to show that $g = \frac{1}{n}\mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi}}[\ln \hat{\mathcal{Z}}_t]$ concentrates w.r.t A_μ (we still keep X_i^* fixed for the moment). To ease the notation we set $\mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi}} = \mathbb{E}_G$ in the rest of this paragraph. We must estimate variations $g - g^{(\nu)}$ corresponding to two vectors $\mathbf{A}$ and $\mathbf{A}^{(\nu)}$ with $A_\mu^{(\nu)} = A_\mu$ for $\mu \neq \nu$ and $A_\nu^{(\nu)} = \tilde{A}_\nu$. By an application of Jensen's inequality one finds

$$\frac{1}{n}\mathbb{E}_G\langle \hat{\mathcal{H}}_t^{(\nu)} - \hat{\mathcal{H}}_t\rangle_{\hat{\mathcal{H}}_t^{(\nu)}} \leq g - g^{(\nu)} \leq \frac{1}{n}\mathbb{E}_G\langle \hat{\mathcal{H}}_t^{(\nu)} - \hat{\mathcal{H}}_t\rangle_{\hat{\mathcal{H}}_t} \tag{127}$$

where the Gibbs brackets pertain to the effective Hamiltonians (122). From (122) we obtain

$$\mathcal{H}_t^{(\nu)} - \mathcal{H}_t = \frac{1}{2}\sum_{\mu=1}^m \big\{\hat{\Gamma}_{t,\mu}^{(\nu)2} - \hat{\Gamma}_{t,\mu}^2 + 2Z_\mu(\hat{\Gamma}_{t,\mu}^{(\nu)} - \hat{\Gamma}_{t,\mu})\big\}$$

Now, a look at equation (115) shows that for $\mu \neq \nu$ we have $(\Gamma_{t,\mu}^{(\nu)})^2 = (\Gamma_{t,\mu})^2$, and therefore only the term $\mu = \nu$ survives in this sum. Consequently

$$\frac{1}{2n}\mathbb{E}_G\Big\langle \hat{\Gamma}_{t,\nu}^{(\nu)2} - \hat{\Gamma}_{t,\nu}^2 + 2Z_\nu(\hat{\Gamma}_{t,\nu}^{(\nu)} - \hat{\Gamma}_{t,\nu})\Big\rangle_{\mathcal{H}_t^{(\nu)}} \leq g - g^{(\nu)} \leq \frac{1}{2n}\mathbb{E}_G\Big\langle \hat{\Gamma}_{t,\nu}^{(\nu)2} - \hat{\Gamma}_{t,\nu}^2 + 2Z_\nu(\hat{\Gamma}_{t,\nu}^{(\nu)} - \hat{\Gamma}_{t,\nu})\Big\rangle_{\mathcal{H}_t}. \tag{128}$$

Using

$$\big|\hat{\Gamma}_{t,\nu}^{(\nu)2} - \Gamma_{t,\nu}^2 + 2Z_\nu(\hat{\Gamma}_{t,\nu}^{(\nu)} - \hat{\Gamma}_{t,\nu})\big| \leq 4\sup|\varphi|^2 + 4|Z_\nu|\sup|\varphi|,$$

from (128) we conclude

$$|g - g^{(\nu)}| \leq \frac{2}{n}\sup|\varphi|(\sup|\varphi| + 2\sqrt{\frac{2}{\pi}}). \tag{129}$$

An application of Proposition H.4 yields

$$\mathbb{P}\Big(\big|\frac{1}{n}\mathbb{E}_G[\ln\hat{\mathcal{Z}}_t] - \frac{1}{n}\mathbb{E}_{G,\mathbf{A}}[\ln\hat{\mathcal{Z}}_t]\big| > r\Big) \leq 2e^{-C(\varphi,S,\alpha)nr^2}. \tag{130}$$

where we recall $\mathbb{E}_G = \mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi}}$ and $\mathbb{P}$ the probability w.r.t all random variables.

C. Bounded difference with respect to X_i^*

Set $\mathbb{E}_\Theta = \mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi},\mathbf{A}}$ for all quenched variables except $\mathbf{X}^*$. We prove concentration of $g = \frac{1}{n}\mathbb{E}_\Theta[\ln\hat{\mathcal{Z}}_t]$ with respect to $\mathbf{X}^*$. This is done by showing a bounded difference property for

$$g - g^{(j)} = \frac{1}{n}\mathbb{E}_\Theta\Bigg[\frac{\hat{\mathcal{Z}}_t}{\hat{\mathcal{Z}}_t^{(j)}}\Bigg]$$

where $\hat{\mathcal{Z}}_t$ and $\hat{\mathcal{Z}}_t^{(j)}$ are the partition functions corresponding to two input signals $\mathbf{X}^*, \mathbf{X}^{*(j)}$ such that $X_i^{*(j)} = X_i$ for $i \neq j$ and $X_j^{*(j)} = \tilde{X}_j^*$. By Jensen's inequality,

$$\begin{aligned}
&\frac{1}{2n}\sum_{\mu=1}^m \mathbb{E}_\Theta\Big\langle \hat{\Gamma}_{t,\mu}^{(j)2} - \hat{\Gamma}_{t,\mu}^2 + 2Z_\mu(\hat{\Gamma}_{t,\mu}^{(j)} - \hat{\Gamma}_{t,\mu})\Big\rangle_{\hat{\mathcal{H}}_t^{(j)}} + \frac{1}{2n}\mathbb{E}_\Theta\Big\langle rt(X_j^* - \tilde{X}_j^*) - 2\sqrt{tr}(x_j + Z_j')(X_j^* - \tilde{X}_j^*)\Big\rangle_{\hat{\mathcal{H}}_t^{(j)}} \\
&\leq g - g^{(j)} \leq \frac{1}{2n}\sum_{\mu=1}^m \mathbb{E}_\Theta\Big\langle \hat{\Gamma}_{t,\mu}^{(j)2} - \hat{\Gamma}_{t,\mu}^2 + 2Z_\mu(\hat{\Gamma}_{t,\mu}^{(j)} - \hat{\Gamma}_{t,\mu})\Big\rangle_{\hat{\mathcal{H}}_t} + \frac{1}{2n}\mathbb{E}_\Theta\Big\langle rt(X_j^* - \tilde{X}_j^*) - 2\sqrt{tr}(x_j + Z_j')(X_j^* - \tilde{X}_j^*)\Big\rangle_{\hat{\mathcal{H}}_t}
\end{aligned} \tag{131}$$

The second expectations on each side of the inequality are obviously bounded by $O(n^{-1})$ for P_0 with bounded support. The other terms are more tedious to treat carefully. We have

$$\begin{aligned}
&\hat{\Gamma}_{t,\mu}^{(j)2} - \hat{\Gamma}_{t,\mu}^2 + 2Z_\mu(\hat{\Gamma}_{t,\mu}^{(j)} - \hat{\Gamma}_{t,\mu}) \\
&= \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2 - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2 \\
&- 2\Big\{\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big) - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{X}^*]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)\Big\} \\
&\times\Big\{\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi}\mathbf{x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) - 2Z_\mu\Big\}
\end{aligned}$$

Taking the Gibbs brackets $\langle - \rangle_{\hat{\mathcal{H}}_t}$ or $\langle - \rangle_{\hat{\mathcal{H}}_t^{(j)}}$ and the expectation $\mathbb{E}_\Theta$ we see that the term Z_μ vanishes (because $\mathbb{E}[Z_\mu] = 0$) and the upper and lower bounds in (131) are both of the form $T_1 + T_2$ with

$$T_1 = \frac{1}{2n} \sum_{\mu=1}^{m} \mathbb{E}_\Theta \Big[\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2 - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2 \Big]$$

$$T_2 = \frac{1}{n} \sum_{\mu=1}^{m} \mathbb{E}_\Theta \Big[\Big\{ \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big) - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big) \Big\}$$
$$\times \Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle \Big]$$

Here we denote by $\langle - \rangle$ either Gibbs bracket $\langle - \rangle_{\hat{\mathcal{H}}_t}$ or $\langle - \rangle_{\hat{\mathcal{H}}_t^{(j)}}$.

1) Estimating T_1: We note that the arguments of φ only differ by $\sqrt{(1-t)/n}\Phi_{\mu j}(\tilde{X}_j^* - X_j^*)$. Thus Taylor expanding each term to second order we find

$$\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2 - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)^2$$
$$= \sqrt{\frac{1-t}{n}}\Phi_{\mu j}(\tilde{X}_j^* - X_j^*)2(\varphi\partial_x\varphi)_{\sim\Phi_{\mu j}} + (\text{remainder})_1$$

where $(\varphi\partial_x\varphi)_{\sim\Phi_{\mu j}}$ mean that the argument of $\varphi\partial\varphi$ is independent of $\Phi_{\mu j}$. It is easy to show (e.g., using a Lagrange formula) $(\text{remainder})_1 \leq C(\varphi, S)n^{-1}\Phi_{\mu j}^2$ because φ bounded and twice differentiable with bounded first and second derivative. When we average over $\Phi_{\mu j}$ the first term disappears and for the second $\mathbb{E}_{\Phi_{\mu j}}[(\text{remainder})_1] \leq C(\varphi, S)n^{-1}$. Therefore we obtain

$$|T_1| \leq C(\varphi, S, \alpha)n^{-1} \tag{132}$$

2) Estimating T_2: Proceeding by a similar Taylor expansion for the difference of φ's in T_2 we find

$$\varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*(j)}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big) - \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi X}^{*}]_\mu + k_1(t)V_\mu + k_2(t)W_\mu^*, A_\mu\Big)$$
$$\times \Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle$$
$$= \sqrt{\frac{1-t}{n}}\Phi_{\mu j}(\tilde{X}_j^* - X_j^*)(\partial_x\varphi)_{\sim\Phi_{\mu j}} \Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle + (\text{remainder})_2 \tag{133}$$

where $(\text{remainder})_2 \leq C(\varphi, S)m^{-1}\Phi_{\mu j}^2$. If we average over $\Phi_{\mu j}$ the first term does not directly disappear because the Gibbs bracket $\langle - \rangle$ stil depends on $\Phi_{\mu j}$ and we have to work a little bit more. We use the mean value theorem to write

$$\Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle = \Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}\sum_{i\neq j}^{n}\Phi_{\mu i}x_i + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle$$
$$+ \Phi_{\mu j}\frac{d}{d\Phi_{\mu j}}\Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}[\mathbf{\Phi x}]_\mu + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle\Big|_{\xi_{\mu j}}$$

for some $0 \leq \xi_{\mu j} \leq \Phi_{\mu j}$. The derivative is of the form

$$\frac{d}{d\Phi_{\mu j}}\langle\varphi\rangle = \sqrt{\frac{1-t}{n}}\langle\partial_x\varphi\rangle + \langle\varphi(\Gamma_{t,\mu} + Z_\mu)\frac{d\hat{\Gamma}_{t,\mu}}{d\Phi_{\mu j}}\rangle - \langle\varphi\rangle\langle(\hat{\Gamma}_{t,\mu} + Z_\mu)\frac{d\hat{\Gamma}_{t,\mu}}{d\Phi_{\mu j}}\rangle$$
$$= \sqrt{\frac{1-t}{n}}\Big\{\langle\partial_x\varphi\rangle + \langle\varphi(\Gamma_{t,\mu} + Z_\mu)(X_j^* - x_j)\partial_x\varphi\rangle - \langle\varphi\rangle\langle(\Gamma_{t,\mu} + Z_\mu)(X_j^* - x_j)\partial_x\varphi\rangle\Big\}$$

where it is understood that $\hat{\Gamma}_{t,\mu}$, X_j^* appears for the bracket $\langle - \rangle_{\hat{\mathcal{H}}_t}$ and $\hat{\Gamma}_{t,\mu}^{(j)}$, $\tilde{X}_j$ appears for $\langle - \rangle_{\hat{\mathcal{H}}_t}$. Putting all these remarks together (133) becomes equal to

$$\sqrt{\frac{1-t}{n}}\Phi_{\mu j}(\tilde{X}_j^* - X_j^*)(\partial_x\varphi)_{\sim\Phi_{\mu j}} \Big\langle \varphi\Big(\sqrt{\frac{1-t}{n}}\sum_{i\neq j}^{n}\Phi_{\mu i}x_i + k_1(t)V_\mu + k_2(t)w_\mu, a_\mu\Big) \Big\rangle + (\text{remainder})_2 + (\text{remainder})_2'$$

where $(\text{remainder})_2' \leq C(\epsilon, \varphi, S)n^{-1}\Phi_{\mu j}^2(1 + |Z_\mu|)$. Now, averaging over $\Phi_{\mu j}$ the first term vanishes, and (with the further average over Z_μ also) both remainders become $O(n^{-1})$. Summarizing, we find

$$|T_2| \leq C(\varphi, S, \alpha)n^{-1}. \tag{134}$$

Finally from (131), (132), (134) we obtain the bounded difference property

$$|g - g^{(j)}| \leq C(\varphi, S, \alpha) n^{-1} \tag{135}$$

and an immediate application of Proposition H.4 then yields

$$\mathbb{P}\Big(\big|\frac{1}{n}\mathbb{E}_\Theta[\ln \hat{\mathcal{Z}}_t] - \frac{1}{n}\mathbb{E}[\ln \hat{\mathcal{Z}}_t]\big| > r\Big) \leq 2e^{-C(\varphi,S,\alpha)r^2 n}. \tag{136}$$

Recall that here $\mathbb{E}_\Theta = \mathbb{E}_{\mathbf{Z},\mathbf{Z}',\mathbf{V},\mathbf{W}^*,\mathbf{\Phi},\mathbf{A}}$ and $\mathbb{P}$ is the probability with respect to all variables $\mathbf{Z}, \mathbf{Z}', \mathbf{V}, \mathbf{W}^*, \mathbf{\Phi}, \mathbf{A}, \mathbf{X}^*$.

D. Proof of Theorem H.1

By the triangle inequality

$$\big|\ln \hat{\mathcal{Z}}_t - \mathbb{E}\ln \hat{\mathcal{Z}}_t\big| \leq \big|\ln \hat{\mathcal{Z}}_t - \mathbb{E}_G \ln \hat{\mathcal{Z}}_t\big| + \big|\mathbb{E}_G \ln \hat{\mathcal{Z}}_t - \mathbb{E}_\Theta \ln \hat{\mathcal{Z}}_t\big| + \big|\mathbb{E}_\Theta \ln \hat{\mathcal{Z}}_t - \mathbb{E}\ln \hat{\mathcal{Z}}_t\big|$$

Therefore by the union bound and and (126), (130) and (136)

$$\begin{aligned}\mathbb{P}\Big(\big|\ln \hat{\mathcal{Z}}_t - \mathbb{E}\ln \hat{\mathcal{Z}}_t\big| > nr\Big) &\leq \mathbb{P}\Big(\big|\ln \hat{\mathcal{Z}}_t - \mathbb{E}_G \ln \hat{\mathcal{Z}}_t\big| > \frac{nr}{3}\Big) + \mathbb{P}\Big(\big|\mathbb{E}_G \ln \hat{\mathcal{Z}}_t - \mathbb{E}_\Theta \ln \hat{\mathcal{Z}}_t\big| > \frac{nr}{3}\Big)\\ &\quad + \mathbb{P}\Big(\big|\mathbb{E}_\Theta \ln \hat{\mathcal{Z}}_t - \mathbb{E}\ln \hat{\mathcal{Z}}_t\big| > \frac{nr}{3}\Big)\\ &\leq 8e^{-\frac{C(\varphi,S,\alpha)r^2 n}{9}}\end{aligned}$$

which is equivalent to (111). To get the second statement (112) of the Theorem we use the observation $\mathbb{E}[X^2] = \int_0^{+\infty} da\, \mathbb{P}(X^2 > a)$ and apply it to $X = n^{-1}\big|\ln \hat{\mathcal{Z}}_t - \mathbb{E}\ln \hat{\mathcal{Z}}_t\big|$.

APPENDIX I
CONCENTRATION OF THE OVERLAP

In this appendix we give the main steps of the proof of Lemma 5.3. We denote by $\langle - \rangle_{n,t,\epsilon}$ the Gibbs measure associated to the perturbed Hamiltonian

$$\mathcal{H}_t(\mathbf{x}, \mathbf{w}; \mathbf{Y}, \mathbf{Y}') + \sum_{i=1}^n \big(\epsilon \frac{x_i^2}{2} - \epsilon x_i X_i^* - \sqrt{\epsilon} x_i \hat{Z}_i\big)$$

i.e., the sum of (62) and (70). It is crucial that the second term is a perturbation which preserves the Nishimori identity of Appendix A. We note that the precise form of the first term does not matter and all subsequent arguments are generic as long as it is a Hamiltonian whose Gibbs distribution satisfies this Nishimori identity. The corresponding average free entropy is denoted $f_{n,\epsilon}(t)$ and we call $F_{n,\epsilon}(t)$ the free entropy for a realisation of the quenched variables, that is $F_{n,\epsilon}(t) = n^{-1}\ln \mathcal{Z}_t(\mathbf{Y}, \mathbf{Y}')$.

Let

$$\mathcal{L}_\epsilon := \frac{1}{n}\sum_{i=1}^n \Big(\frac{x_i^2}{2} - x_i s_i - \frac{x_i \widehat{z}_i}{2\sqrt{\epsilon}}\Big).$$

Up to the prefactor n^{-1} this quantity is the derivative of the perturbation term in (70). The fluctuations of the overlap $Q = n^{-1}\sum_{i=1}^n X_i^* x_i$ and those of $\mathcal{L}_\epsilon$ are related through the remarkable identity

$$\mathbb{E}\big\langle(\mathcal{L}_\epsilon - \mathbb{E}\langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon})^2\big\rangle_{n,t,\epsilon} = \frac{1}{4}\mathbb{E}\langle(Q - \mathbb{E}\langle Q\rangle_{n,t,\epsilon})^2\big\rangle_{n,t,\epsilon} + \frac{1}{2}\mathbb{E}[\langle Q^2\rangle_{n,t,\epsilon} - \langle Q\rangle_{n,t,\epsilon}^2] + \frac{1}{4n^2\epsilon}\sum_{i=1}^n \mathbb{E}[\langle X_i^2\rangle_{n,t,\epsilon} - \langle X_i\rangle_{n,t,\epsilon}^2].$$

A detailed derivation is found in Appendix IX of [46] and involves only some lengthy algebra using the Nishimori identity and integrations by parts w.r.t the Gaussian $\hat{Z}_i$ in the perturbation term. Lemma 5.3 is then a direct consequence of the following:

***Proposition I.1** (Concentration of $\mathcal{L}_\epsilon$ on $\mathbb{E}\,[\langle \mathcal{L}_\epsilon\rangle]$):* Let P_0 with bounded support in $[-S, S]$. For any $0 < a < b < 1$,

$$\lim_{n\to+\infty} \int_a^b d\epsilon\, \mathbb{E}\big\langle(\mathcal{L}_\epsilon - \mathbb{E}\langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon})^2\big\rangle_{n,t,\epsilon} = 0. \tag{137}$$

The proof of this proposition is broken in two parts. Notice that

$$\mathbb{E}\big\langle(\mathcal{L}_\epsilon - \mathbb{E}\langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon})^2\big\rangle_{n,t,\epsilon} = \mathbb{E}\big\langle(\mathcal{L}_\epsilon - \langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon})^2\big\rangle_{n,t,\epsilon} + \mathbb{E}\big[(\langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon} - \mathbb{E}\langle \mathcal{L}_\epsilon\rangle_{n,t,\epsilon})^2\big]. \tag{138}$$

Thus it suffices to prove the two following lemmas. The first lemma expresses concentration w.r.t the posterior distribution (or "thermal fluctuations") and is an elementary consequence of concavity properties of the free energy and the Nishimori identity.

Lemma I.2 (Concentration of $\mathcal{L}_\epsilon$ on $\langle \mathcal{L}_\epsilon \rangle$): Let P_0 with bounded second moment. For any $0 < a < b < 1$ we have

$$\lim_{n\to+\infty} \int_a^b d\epsilon\, \mathbb{E}\big[\big\langle (\mathcal{L}_\epsilon - \langle \mathcal{L}_\epsilon \rangle_{n,t,\epsilon})^2 \big\rangle_{n,t,\epsilon}\big] = 0 \tag{139}$$

The second lemma expresses the concentration of the average overlap w.r.t the realizations of quenched disorder variables and is a consequence of the concentration of the free energy (more precisely equation (112) in Appendix H).

Lemma I.3 (Concentration of $\langle \mathcal{L}_\epsilon \rangle$ on $\mathbb{E}\langle \mathcal{L}_\epsilon \rangle$): Let P_0 with bounded support in $[-S, S]$. For any $0 < a < b < 1$ we have

$$\lim_{n\to+\infty} \int_a^b d\epsilon\, \mathbb{E}\big[(\langle \mathcal{L}_\epsilon \rangle_{n,t,\epsilon} - \mathbb{E}[\langle \mathcal{L}_\epsilon \rangle_{n,t,\epsilon}])^2\big] = 0 \tag{140}$$

The reader is refered to Sec. V of [46] for the proof of these two Lemmas. We point out that the analysis gives a rate of decay $O(n^{-1})$ for (140) which is optimal but a weaker decay rate for (140). However any decay rate will do for the present proof of the replica formula.

Appendix J
Computing the optimal generalization error

A. Generalization error at finite $\Delta^{\rm te}$

Let $w \sim \mathcal{N}(0,1)$ and $a \sim P_A$. It is convenient to introduce the following function

$$f^{\rm te}(y|\sqrt{q}\,V, \rho - q) := \mathbb{E}_w P_{\rm out}^{\rm te}(y|\sqrt{q}\,V + \sqrt{\rho - q}\,w) = \mathbb{E}_{w,a} \frac{e^{-\frac{1}{2\Delta^{\rm te}}\left(y - \varphi(\sqrt{q}\,V + \sqrt{\rho-q}\,w, a)\right)^2}}{\sqrt{2\pi\Delta^{\rm te}}}\,. \tag{141}$$

It is related to the expression of $\Psi_{P_{\rm out}^{\rm te}}(q;\rho)$ that appears in the potential (40). Using (16) we obtain

$$\Psi_{P_{\rm out}^{\rm te}}(q;\rho) = \mathbb{E} \int dy f^{\rm te}(y|\sqrt{q}\,V, \rho - q) \ln f^{\rm te}(y|\sqrt{q}\,V, \rho - q)\,. \tag{142}$$

Denote (q^*, r^*) the values corresponding to $\sup_{q\in[0,\rho]} \inf_{r\geq 0} f_{\rm RS}^{\rm ts}(q, r)$, where $f_{\rm RS}^{\rm ts}$ is the expression inside the brackets in (40). Then the envelope theorem [56] allows to write

$$\frac{d f_{\rm RS}^{\rm ts}(q^*, r^*)}{d(\Delta^{\rm te})^{-1}} = \frac{\partial f_{\rm RS}^{\rm ts}(q, r)}{\partial(\Delta^{\rm te})^{-1}}\Big|_{q^*, r^*}\,. \tag{143}$$

Using the expression of $f_{\rm RS}^{\rm ts}$ and noticing that only the third term $\Psi_{P_{\rm out}^{\rm te}}$ given by (142) depends explicitly on $\Delta^{\rm te}$, one gets (using the dominated convergence theorem)

$$\begin{aligned}
\frac{\partial f_{\rm RS}^{\rm ts}(q, r)}{\partial(\Delta^{\rm te})^{-1}} &= \alpha(1-\beta) \frac{\partial \Psi_{P_{\rm out}^{\rm te}}(q;\rho)}{\partial(\Delta^{\rm te})^{-1}} \\
&= \alpha(1-\beta)\mathbb{E} \int dy \frac{\partial f^{\rm te}(y|\sqrt{q}\,V, \rho - q)}{\partial(\Delta^{\rm te})^{-1}} \big(1 + \ln f^{\rm te}(y|\sqrt{q}\,V, \rho - q)\big) \\
&= \alpha(1-\beta)\mathbb{E} \int dy \frac{\partial f^{\rm te}(y|\sqrt{q}\,V, \rho - q)}{\partial(\Delta^{\rm te})^{-1}} \ln f^{\rm te}(y|\sqrt{q}\,V, \rho - q)\,,
\end{aligned} \tag{144}$$

using that $f^{\rm te}(y|\sqrt{q}\,V, \rho - q)$ is a probability density for the last equality. Another convenient equivalent expression is

$$\frac{\partial f_{\rm RS}^{\rm ts}(q, r)}{\partial(\Delta^{\rm te})^{-1}} = \alpha(1-\beta)\frac{\Delta^{\rm te}}{2} \int dY \frac{\left\{\frac{d}{dy} f^{\rm te}(y|\sqrt{q}\,V, \rho - q)\right\}^2}{f^{\rm te}(y|\sqrt{q}\,V, \rho - q)}\,. \tag{145}$$

The expression (145) is obtained from (144) using the easily checkable identity

$$\frac{\partial f^{\rm te}(y|\sqrt{q}\,V, \rho - q)}{\partial(\Delta^{\rm te})^{-1}} = -\frac{\Delta^{\rm te}}{2} \frac{d^2 f^{\rm te}(y|\sqrt{q}\,V, \rho - q)}{dy^2} \tag{146}$$

and an integration by part. Finally combining (143), (145) and (46), it leads the expression of the generalization error

$$\begin{aligned}
\lim_{n\to\infty} \mathcal{E}_{\rm gen} &= \Delta^{\rm te}\Big(1 - \Delta^{\rm te}\mathbb{E} \int dy \frac{\left\{\frac{d}{dy} f^{\rm te}(y|\sqrt{q^*}\,V, \rho - q^*)\right\}^2}{f^{\rm te}(y|\sqrt{q^*}\,V, \rho - q^*)}\Big) \\
&= \Delta^{\rm te}\Big(1 - \mathbb{E} \int dy \frac{\mathbb{E}_{w,a}\Big[\frac{1}{\sqrt{2\pi\Delta^{\rm te}}} e^{-\frac{1}{2\Delta^{\rm te}}\left(y - \varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)\right)^2} \big(y - \varphi(\sqrt{q^*}\,V + \sqrt{\rho - q^*}\,w, a)\big)\Big]^2}{\mathbb{E}_{w,a}\Big[\frac{1}{\sqrt{2\pi\Delta^{\rm te}}} e^{-\frac{1}{2\Delta^{\rm te}}\left(y - \varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)\right)^2}\Big]}\Big)\,.
\end{aligned} \tag{147}$$

Do not get confused: $\mathbb{E}[\cdot]^2$ is an expectation to the square, and not *of* a squared quantity. For illustrating purpose, simple algebra leads that this formula for the binary perceptron $\varphi(x) = \mathrm{sgn}(x)$ is given by

$$\lim_{n\to\infty} \mathcal{E}_{\rm gen} = \frac{\Delta^{\rm te}}{2}\,\mathbb{E}\Big[(1+u)\int dy \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\big(y^2-1\big)\ln\Big\{\cosh\Big(\frac{y}{\sqrt{\Delta^{\rm te}}}+\frac{1}{\Delta^{\rm te}}\Big)+u\sinh\Big(\frac{y}{\sqrt{\Delta^{\rm te}}}+\frac{1}{\Delta^{\rm te}}\Big)\Big\}\Big]$$
$$+\frac{\Delta^{\rm te}}{2}\,\mathbb{E}\Big[(1-u)\int dy \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\big(y^2-1\big)\ln\Big\{\cosh\Big(\frac{y}{\sqrt{\Delta^{\rm te}}}-\frac{1}{\Delta^{\rm te}}\Big)+u\sinh\Big(\frac{y}{\sqrt{\Delta^{\rm te}}}-\frac{1}{\Delta^{\rm te}}\Big)\Big\}\Big] \tag{148}$$

where $u = u(V,q,\rho) := \mathrm{erf}(V\sqrt{q/(2(\rho-q))})$. One can also check directly from this formula instead of the general one (38) that (50) is recovered in the high test noise $\Delta^{\rm te} \to \infty$ limit.

Let mention that expanding the square in (147), one directly obtains that this expression of the generalization error at finite noise $\Delta^{\rm te}$ in the test set matches the equivalent expression (25).

B. Taking the $\Delta^{\rm te} \to \infty$ limit

Let us now consider the high noise limit of (147). Let us denote the integral appearing in it as I. Denote for this subsection $\varphi_* := \varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)$ and let $y \sim \mathcal{N}(0,1)$. First, using a change of variable and isolating a Gaussian probability density function, we rewrite it as

$$I = \Delta^{\rm te}\mathbb{E}_y \frac{\mathbb{E}_{w,a}\Big[e^{\frac{y\varphi_*}{\sqrt{\Delta^{\rm te}}}-\frac{\varphi_*^2}{2\Delta^{\rm te}}}\big(y-\frac{\varphi_*}{\sqrt{\Delta^{\rm te}}}\big)\Big]^2}{\mathbb{E}_{w,a}e^{\frac{y\varphi_*}{\sqrt{\Delta^{\rm te}}}-\frac{\varphi_*^2}{2\Delta^{\rm te}}}} = \Delta^{\rm te}\mathbb{E}_y \frac{\mathbb{E}_{w,a}\Big[\big(1+\frac{y\varphi_*}{\sqrt{\Delta^{\rm te}}}-\frac{\varphi_*^2}{2\Delta^{\rm te}}+\frac{y^2\varphi_*^2}{2\Delta^{\rm te}}\big)\big(y-\frac{\varphi_*}{\sqrt{\Delta^{\rm te}}}\big)\Big]^2}{\mathbb{E}_{w,a}\big[1+\frac{y\varphi_*}{\sqrt{\Delta^{\rm te}}}-\frac{\varphi_*^2}{2\Delta^{\rm te}}+\frac{y^2\varphi_*^2}{2\Delta^{\rm te}}\big]} + \mathcal{O}((\Delta^{\rm te})^{-1})\,. \tag{149}$$

Now denote $\varphi_1 := \mathbb{E}_{w,a}\varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)$ and $\varphi_2 := \mathbb{E}_{w,a}[\varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)^2]$. Expanding around $(\Delta^{\rm te})^{-1} \to 0$ we get

$$I = \Delta^{\rm te}\mathbb{E}_y\Big[y^2 + \frac{\varphi_1(y^3-2y)}{\sqrt{\Delta^{\rm te}}} + \frac{1}{\Delta^{\rm te}}\Big(\varphi_1^2 - \frac{5}{2}\varphi_2 y^2 + \frac{1}{2}\varphi_2 y^4\Big)\Big] + \mathcal{O}((\Delta^{\rm te})^{-1}) = \Delta^{\rm te} + \varphi_1^2 - \varphi_2 + \mathcal{O}((\Delta^{\rm te})^{-1})\,. \tag{150}$$

Plugging this in (147) finally leads $\lim_{n\to\infty}\mathcal{E}_{\rm gen} = \mathbb{E}_V[\varphi_2 - \varphi_1^2] + \mathcal{O}((\Delta^{\rm te})^{-1})$. The last step for obtaining (38) is to notice that $\mathbb{E}_V\varphi_2 = \mathbb{E}_{V,w,a}\big[\varphi(\sqrt{q^*}\,V + \sqrt{\rho-q^*}\,w, a)^2\big] = \mathbb{E}_{V,a}\big[\varphi(\sqrt{\rho}\,V, a)^2\big]$ as V and w are i.i.d $\mathcal{N}(0,1)$ random variables.

Acknowledgments

Jean Barbier acknowledges funding from the Swiss National Science Foundation (grant 200021-156672). Florent Krzakala acknowledges funding from the ERC under the European Union 7th Framework Programme Grant Agreement 307087-SPARCS. Part of this work was done while Léo Miolane was visiting EPFL.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[2] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
[3] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18 914–18 919, 2009.
[4] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, Dec 2006.
[5] C. E. Shannon, "A mathematical theory of communication, part i, part ii," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, 1948.
[6] M. Mézard, G. Parisi, and M.-A. Virasoro, "Spin glass theory and beyond." 1987.
[7] H. S. Seung, H. Sompolinsky, and N. Tishby, "Statistical mechanics of learning from examples," *Phys. Rev. A*, vol. 45, pp. 6056–6091, Apr 1992. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.45.6056
[8] T. L. H. Watkin, A. Rau, and M. Biehl, "The statistical mechanics of learning a rule," *Rev. Mod. Phys.*, vol. 65, pp. 499–556, Apr 1993. [Online]. Available: https://link.aps.org/doi/10.1103/RevModPhys.65.499
[9] K. H. Fischer and J. A. Hertz, *Spin glasses*. Cambridge university press, 1993, vol. 1.
[10] V. Dotsenko, *An introduction to the theory of spin glasses and neural networks*. World Scientific, 1995, vol. 54.
[11] A. Engel and C. Van den Broeck, *Statistical mechanics of learning*. Cambridge University Press, 2001.
[12] H. Nishimori, *Statistical physics of spin glasses and information processing: an introduction*. Clarendon Press, 2001, vol. 111.
[13] M. Mezard and A. Montanari, *Information, physics, and computation*. Oxford University Press, 2009.
[14] L. Zdeborová and F. Krzakala, "Statistical physics of inference: thresholds and algorithms," *Advances in Physics*, vol. 65, no. 5, pp. 453–552, 2016.
[15] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, no. 36, pp. 14 557–14 562, 2013.
[16] M. Bayati and A. Montanari, "The lasso risk for gaussian matrices," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 1997–2017, April 2012.
[17] D. Donoho and A. Montanari, "High dimensional robust m-estimation: asymptotic variance via approximate message passing," *Probability Theory and Related Fields*, vol. 166, no. 3, pp. 935–969, Dec 2016. [Online]. Available: https://doi.org/10.1007/s00440-015-0675-z
[18] R. Gribonval and P. Machart, "Reconciling" priors" &" priors" without prejudice?" in *Advances in Neural Information Processing Systems*, 2013, pp. 2193–2201.
[19] M. Advani and S. Ganguli, "An equivalence between high dimensional bayes optimal inference and m-estimation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3378–3386.

[20] D. Donoho and J. Tanner, “Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing,” *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4273–4293, 2009.

[21] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, Nov 2009.

[22] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *2011 IEEE International Symposium on Information Theory Proceedings*, July 2011, pp. 2168–2172.

[23] P. T. Boufounos and R. G. Baraniuk, “1-bit compressive sensing,” in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*. IEEE, 2008, pp. 16–21.

[24] J. Barbier, “Statistical physics and approximate message-passing algorithms for sparse linear estimation problems in signal processing and coding theory,” Ph.D. dissertation, Université Paris Diderot, 2015. [Online]. Available: http://arxiv.org/abs/1511.01650

[25] J. Barbier, M. Dia, N. Macris, F. Krzakala, T. Lesieur, and L. Zdeborová, “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula,” in *Advances in Neural Information Processing Systems 29*, 2016, p. 424–432.

[26] M. Lelarge and L. Miolane, “Fundamental limits of symmetric low-rank matrix estimation,” *ArXiv e-prints*, Nov. 2016.

[27] L. Miolane, “Fundamental limits of low-rank matrix estimation: the non-symmetric case,” *ArXiv e-prints*, Feb. 2017.

[28] T. Lesieur, L. Miolane, M. Lelarge, F. Krzakala, and L. Zdeborová, “Statistical and computational phase transitions in spiked tensor estimation,” *ArXiv e-prints*, Jan. 2017.

[29] T. Tanaka, “A statistical-mechanics approach to large-system analysis of cdma multiuser detectors,” *IEEE Transactions on Information Theory*, vol. 48, no. 11, pp. 2888–2910, Nov 2002.

[30] D. Guo and S. Verdú, “Randomly spread cdma: Asymptotics via statistical physics,” *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1983–2010, June 2005.

[31] A. R. Barron and A. Joseph, “Toward fast reliable communication at rates near capacity with gaussian noise,” in *2010 IEEE International Symposium on Information Theory*, June 2010, pp. 315–319.

[32] J. Barbier and F. Krzakala, “Approximate message-passing decoder and capacity-achieving sparse superposition codes,” 2015. [Online]. Available: http://arxiv.org/abs/1503.08040

[33] J. Barbier, M. Dia, and N. Macris, “Proof of threshold saturation for spatially coupled sparse superposition codes,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 1173–1177.

[34] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, “Inference and Phase Transitions in the Detection of Modules in Sparse Networks,” *Physical Review Letters*, vol. 107, no. 6, p. 065701, Aug. 2011.

[35] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, “Spectral redemption in clustering sparse networks,” *Proceedings of the National Academy of Science*, vol. 110, pp. 20 935–20 940, Dec. 2013.

[36] F. Caltagirone, M. Lelarge, and L. Miolane, “Recovering asymmetric communities in the stochastic block model,” *ArXiv e-prints*, Oct. 2016.

[37] E. Abbe, “Community detection and stochastic block models: recent developments,” *ArXiv e-prints*, Mar. 2017.

[38] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, “Gibbs states and the set of solutions of random constraint satisfaction problems,” *Proceedings of the National Academy of Science*, vol. 104, pp. 10 318–10 323, Jun. 2007.

[39] J. Barbier, F. Krzakala, L. Zdeborová, and P. Zhang, “The hard-core model on random graphs revisited,” in *Journal of Physics Conference Series*, ser. Journal of Physics Conference Series, vol. 473, Dec. 2013, p. 012021.

[40] C. Baldassi, A. Braunstein, N. Brunel, and R. Zecchina, “Efficient supervised learning in networks with binary synapses,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 26, pp. 11 079–11 084, 2007.

[41] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, “Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7655–E7662, 2016. [Online]. Available: http://www.pnas.org/content/113/48/E7655.abstract

[42] J. A. Nelder and R. J. Baker, *Generalized linear models*. Wiley Online Library, 1972.

[43] E. Gardner and B. Derrida, “Three unfinished works on the optimal storage capacity of networks,” *Journal of Physics A: Mathematical and General*, vol. 22, no. 12, p. 1983, 1989.

[44] G. Györgyi, “First-order transition to perfect generalization in a neural network with binary synapses,” *Physical Review A*, vol. 41, no. 12, p. 7097, 1990.

[45] E. B. Baum and Y.-D. Lyuu, “The transition to perfect generalization in perceptrons,” *Neural computation*, vol. 3, no. 3, pp. 386–401, 1991.

[46] J. Barbier and N. Macris, “The stochastic interpolation method: A simple scheme to prove replica formulas in bayesian inference,” *CoRR*, vol. abs/1705.02780, 2017. [Online]. Available: http://arxiv.org/abs/1705.02780

[47] J. Barbier, M. Dia, N. Macris, and F. Krzakala, “The mutual information in random linear estimation,” in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2016.

[48] J. Barbier, N. Macris, M. Dia, and F. Krzakala, “Mutual Information and Optimality of Approximate Message-Passing in Random Linear Estimation.” [Online]. Available: https://arxiv.org/pdf/1701.05823v1.pdf

[49] G. Reeves and H. D. Pfister, “The replica-symmetric prediction for compressed sensing with gaussian matrices is exact,” in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 665–669.

[50] M. Opper and D. Haussler, “Generalization performance of bayes optimal classification algorithm for learning a perceptron,” *Physical Review Letters*, vol. 66, no. 20, p. 2677, 1991.

[51] D. J. Thouless, P. W. Anderson, and R. G. Palmer, “Solution of‘solvable model of a spin glass’,” *Philosophical Magazine*, vol. 35, no. 3, p. 593–601, 1977.

[52] M. Mézard, “The space of interactions in neural networks: Gardner’s computation with the cavity method,” *Journal of Physics A: Mathematical and General*, vol. 22, no. 12, pp. 2181–2190, 1989.

[53] Y. Kabashima, “Inference from correlated patterns: a unified theory for perceptron learning and linear vector channels,” *Journal of Physics: Conference Series*, vol. 95, no. 1, p. 012001, 2008. [Online]. Available: http://stacks.iop.org/1742-6596/95/i=1/a=012001

[54] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, “Statistical-physics-based reconstruction in compressed sensing,” *Phys. Rev. X*, vol. 2, p. 021005(18), May 2012.

[55] J. P. Vila and P. Schniter, “Expectation-maximization gaussian-mixture approximate message passing,” *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4658–4672, 2013.

[56] P. Milgrom and I. Segal, “Envelope theorems for arbitrary choice sets,” *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

[57] T. Richardson and R. Urbanke, *Modern coding theory*. Cambridge university press, 2008.

[58] M. J. Wainwright, M. I. Jordan *et al.*, “Graphical models, exponential families, and variational inference,” *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[59] J. Ziniel, P. Schniter, and P. Sederberg, “Binary linear classification and feature selection via generalized approximate message passing,” in *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*. IEEE, 2014, pp. 1–6.

[60] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, Feb 2011.

[61] M. Bayati, M. Lelarge, and A. Montanari, “Universality in polytope phase transitions and message passing algorithms,” *The Annals of Applied Probability*, vol. 25, no. 2, pp. 753–822, 2015.

[62] 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*, vol. 2, no. 2, pp. 115–144, 2013.

[63] D. Guo, S. Shamai, and S. Verdú, “Mutual information and minimum mean-square error in gaussian channels,” *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

[64] Y. Xu, Y. Kabashima, and L. Zdeborová, “Bayesian signal reconstruction for 1-bit compressed sensing,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 11, p. P11015, 2014.

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[66] F. Chollet, “keras,” https://github.com/fchollet/keras, 2015.

[67] F. Guerra and F. L. Toninelli, “The thermodynamic limit in mean field spin glass models,” *Communications in Mathematical Physics*, vol. 230, no. 1, pp. 71–79, 2002.

[68] M. Talagrand, *Mean field models for spin glasses: Volume I: Basic examples*. Springer Science & Business Media, 2010, vol. 54.

[69] N. Macris, “Griffith-kelly-sherman correlation inequalities: A useful tool in the theory of error correcting codes,” *IEEE Transactions on Information Theory*, vol. 53, no. 2, pp. 664–683, Feb 2007.

[70] S. B. Korada and N. Macris, “Tight bounds on the capacity of binary input random cdma systems,” *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5590–5613, Nov 2010.

[71] ——, “Exact solution of the gauge symmetric p-spin glass model on a complete graph,” *Journal of Statistical Physics*, vol. 136, no. 2, pp. 205–230, 2009.