Mutual Information in Rank-One Matrix Estimation

Florent Krzakala*,†§, Jiaming Xu§ and Lenka Zdeborová†§

* Laboratoire de Physique Statistique, Sorbonnes Universités & Université Pierre & Marie Curie, 75005 Paris
† Laboratoire de Physique Statistique (CNRS UMR-8550), PSL Universités & École Normale Supérieure, 75005 Paris
§ Simons Institute for the Theory of Computing, University of California, Berkeley, Berkeley, CA, 94720
‡ Institut de Physique Théorique, CNRS, CEA, Université Paris-Saclay, F-91191, Gif-sur-Yvette, France.

Abstract—We consider the estimation of a \(n\)-dimensional vector \(x\) from the knowledge of noisy and possibly non-linear element-wise measurements of \(xx^T\), a very generic problem that contains, e.g., stochastic 2-block model, submatrix localization or the spike perturbation of random matrices. We use an interpolation method proposed by Guerra [1] and later refined by Korada and Macris [2]. We prove that the Bethe mutual information (related to the Bethe free energy and conjectured to be exact by Lesieur et al. [3] on the basis of the non-rigorous cavity method) always yields an upper bound to the exact mutual information. We also provide a lower bound using a similar technique. For concreteness, we illustrate our findings on the sparse PCA problem, and observe that (a) our bounds match for a large region of parameters and (b) that it exists a phase transition in a region where the spectrum remains uninformative. While we present only the case of rank-one symmetric matrix estimation, our proof technique is readily extendable to low-rank symmetric matrix or low-rank symmetric tensor estimation.

I. INTRODUCTION AND MAIN RESULTS

The estimation of low-rank matrices from their noisy, incomplete, or non linear measurements is a problem that has a wide range of applications of practical interest in machine learning and statistics, ranging from the sparse PCA [4], and community detection [5] to sub-matrix localization [6, 7]. We shall consider the setting where the rank-one matrix to be estimated is created as:

\[
W = \frac{1}{\sqrt{n}} xx^T,
\]

where \(x\) is a \(n\)-dimensional vector whose elements were chosen independently at random from a prior distribution \(p(x)\). The matrix \(W\) is then observed element-wise through a noisy non-linear output channel \(P_{\text{out}}(y_{ij} | W_{ij})\), with \(i, j = 1, \ldots, n\). We assume the noise to be symmetric so that \(Y_{ij} = Y_{ji}\). The goal is to estimate the unknown vector \(x\) from \(Y\) up to a global flip of sign, or equivalently the unknown rank-one matrix \(W\) from \(Y\). Throughout this paper, we assume that \(p(x)\) and \(P_{\text{out}}(y, w)\) are independent of \(n\), and are known.

We consider here an information-theoretic viewpoint and analyze the mutual information for the above model defined as \(I(W; Y) = I(x; Y) = E_{x, Y} \log [P(x, Y)/p(x)P(Y)]\). Up to a simple term, see eq. (10), the mutual information is related to the free energy, which is the fundamental quantity usually considered in statistical physics [8, 9]. Recently, an explicit single-letter characterization of the mutual information between the noisy observation and the vector to be recovered has been computed in some special cases of our setting [4, 5]. The general formula has been derived by Lesieur et al. [3, 10] on the basis of the heuristic cavity method from statistical mechanics [8, 9, 11]. We shall refer to the formula conjectured in [3] as the Bethe mutual information. In this contribution, we use a rigorous technique that also originated in physics, the so-called Guerra interpolation [1, 2], to prove that the Bethe mutual information provides always an upper bound. By a variant of the Guerra interpolation, we also provide a lower bound on the mutual information that matches the upper bound for a sizable range of parameters.

A. Main results

Our first result is a rigorous proof of a conjecture from Lesieur et al. [3] of channel universality. In the context of community detection in graphs with growing average degrees, an equivalence between Bernoulli channel and Gaussian channel has been proven already in [5].

Theorem I.1 (Channel Universality). Assume model (1) with a prior \(p(x)\) having a finite support, and the output channel \(P_{\text{out}}(y|w)\) such that at \(w = 0, \log P_{\text{out}}(y|w)\) is thrice differentiable with bounded second and third derivatives and \(\mathbb{E}_{P_{\text{out}}(y|0)}[\frac{\partial}{\partial w} \log P_{\text{out}}(y|w)|_{w=0}] = O(1)\). Then the mutual information per variable satisfies

\[
I(W; Y) = I(W; W + \sqrt{\Delta} \xi) + O(\sqrt{n}),
\]

where \(\xi\) is a symmetric matrix such that \(\xi_{i,j} \sim N(0, 1)\) for \(i \leq j\), and \(\Delta\) is the inverse Fisher information (evaluated at \(w = 0\)) of the channel \(P_{\text{out}}(y|w)\):

\[
\frac{1}{\Delta} = \mathbb{E}_{P_{\text{out}}(y|0)} \left[ (\frac{\partial \log P_{\text{out}}(y|w)}{\partial w} |_{y,0})^2 \right].
\]

Informally, this means that we only have to compute the mutual information for a Gaussian channel to take care of a wide range of channels.

Our next result is that, the Bethe mutual information is always an upper bound to the true one for any finite \(n\):

Theorem I.2 (Upper Bound). Assume model (1) with a prior \(p(x)\) having finite support, and a Gaussian channel such that \(P_{\text{out}}(y|w)\) is the probability density function of a centered Gaussian distribution with variance \(\Delta\). Then for all non-negative parameter \(m\), the mutual information per variable \(I(x; Y)/n\) is upper bounded by the Bethe mutual information

\[
I(x; Y)/n \leq I(W; Y) + O(1).
\]
where $\mathbb{E}_x$ denotes the expectation taken over the prior distribution $p(x)$, $z$ is a Gaussian variable following $\mathcal{N}(0,1)$, and where

$$\mathcal{J}(A, B) = \log \int e^{Bx - Ax^2/2}p(x)dx.$$  \hspace{1cm} (5)

Note that the mutual information is related to the free energy only by a simple term eq. (10).

Our last main result yields an asymptotic lower bound:

**Theorem I.3 (Lower Bound).** With the same hypothesis as in Theorem I.2, denote $\hat{m}$ the minimizer of eq. (4). Define

$$i_L(m) = \frac{2m^2 - \hat{m}^2 + \mathbb{E}_x(x^2)}{4\Delta} - \mathbb{E}_{x,z} \left[ \mathcal{J} \left( \hat{m}, \frac{mx}{\Delta} + \sqrt{\hat{m}z} \right) \right],$$  \hspace{1cm} (6)

where $z \sim \mathcal{N}(0,1)$. Assume that $i_L(m) = 0$ has a finite number of solutions. Then

$$\lim inf_{n \to \infty} \frac{1}{n} I(x; Y) \geq \min_m i_L(m).$$

One can verify that $\hat{m}$ is always a stationary point of $i_L(m)$. If additionally $\hat{m} = \tilde{m}$, where $\tilde{m} = \arg \min_m i_L(m)$, then $\min_m i_L(m) = i_L(\tilde{m})$ and the Bethe mutual information asymptotically equals the true one. As we shall see this is the case for some range of parameters, but not always.

**B. Relation to previous works**

For two particular cases of model (1), the mutual information has previously been proven rigorously [4, 5] using the approximate message passing algorithm and its state evolution [12]. Remarkably, these were constructive proofs, with an explicit algorithm that achieves the minimum mean squared error (MMSE). The proof technique of [4, 5] does not extend straightforwardly when the state evolution had more than one fixed point. Our approach applies to more general class of problems (even when several fixed points exist), but is not constructive and our lower bound is not always tight.

We rely on two essential contributions. First, we use the cavity computations of Lesieur et al. who solved the problem using statistical physics methods [3]. Our results are a considerable step towards confirming the full validity of this approach. We show that the Bethe mutual information always yields an upper bound, and by our lower bound we confirm that the Bethe mutual information is exact for a large range of parameters. Secondly, our approach is inspired by the scheme introduced by Korada and Macris [2] for studying the Sherrington-Kirkpatrick model of spin glasses on the so-called Nishimori line with the Guerra interpolation [1]. It also crucially exploits the Nishimori identities [8] for optimal Bayesian estimation [13–15]. It is worth to remark that for problems of Bayes-optimal estimation (on the Nishimori line) the simplest version of the Guerra interpolation provides an upper bound on the free energy/mutual information while for standard statistical mechanics models, or for optimization problems, it provides instead a lower bound [1, 16].

While we present only the rank-one version on model (1), our proof is readily extendable to any finite rank, or even to tensor factorization. Future directions include the extension to non-symmetric matrices, which are less straightforward. We believe that our results, once more, give strong credibility to the use of the replica and the cavity methods for statistical estimation problems.

**II. APPLICATION TO SPARSE RADEMACHER VARIABLES**

We shall illustrate (see Fig. 1), for concreteness, our findings on a specific example of sparse Rademacher variables where $x = 0, +1, -1$ with probability $1 - \rho$, $\rho / 2$, $\rho / 2$ respectively.

Solving numerically for (4) and (6) shows that the region where the two bounds coincide, and the Bethe mutual information is thus rigorously exact, is quite sizable. This happens, for instance, for all $\rho$ if $\Delta$ is larger than a value around 0.15, and for all $\Delta$ if $\rho$ is larger than a value around 0.66. For large enough $\Delta$, we also find that $\hat{m} = \tilde{m} = 0$, and a derivative of the mutual information indicates that it is impossible to find an assignment correlated to the truth.

In the non-sparse case ($\rho = 1$), one can further show that the minimizers of both (4) and (6) always coincide and the Bethe mutual information is thus asymptotically exact for all $\Delta > 0$. A phase transition arises at $\Delta = 1$ so that for $\Delta < 1$, a non-trivial solution with $\hat{m} = \tilde{m} > 0$ appears. Among the problems
that belong into this category are the dense version of the binary stochastic block model [5,17,18], the dense version of the censored block model [19–21], or the Sherrington-Kirkpatrick model on the Nishimori line as originally studied by Korada and Macris [2].

When both $\rho$ and $\Delta$ are small enough, numerically we find that our upper and lower bounds stop to coincide. Define $\Delta_{\text{Match}}(\rho)$ as the minimum $\Delta$ so that our upper and lower bounds match, for a fixed $\rho$. This is illustrated in Fig. 1 with three different values of $\rho$. When $\Delta < \Delta_{\text{Algo}} = \sqrt{\rho}$, polynomial-time algorithms such as message passing [3,5,10,12] or spectral methods [22] are known to be able to find an assignment with a non-trivial correlation to the truth; thus, in this region, the non-trivial detection is easy. We define $\Delta_{\text{Detect}}$ to be the (conjectured) information-theoretic (IT) threshold for the non-trivial detection. Depending on the particular values of $\rho$, we have the following two strikingly different observations:

(1) When $\rho$ is large (e.g. $\rho = 0.6$), $\Delta_{\text{Match}} < \Delta_{\text{Algo}} = \Delta_{\text{Detect}}$. In this case, our upper and lower bound coincide, showing that there is a non-analyticity (phase transition) at $\Delta_{\text{Detect}}$. In this region, the conjectured IT threshold is indeed the true one and coincides with the algorithmic one $\Delta_{\text{Algo}}$.

(2) When $\rho$ becomes smaller than a certain threshold $\rho^*$, $\Delta_{\text{Algo}} < \Delta_{\text{Detect}} < \Delta_{\text{Match}}$. Numerically we find that $\rho^* \approx 0.09$. The derivative of $i_{jB}$ undergoes a phase transition at $\Delta_{\text{Detect}}$. It readily implies that the derivative of the true mutual information per variable must exhibit a phase transition somewhere between $\Delta_{\text{Match}}$ and $\Delta_{\text{Detect}}$, which is strictly above $\Delta_{\text{Algo}}$. Hence, in this case, there exists a region where the non-trivial detection is informationally possible, but it is impossible via standard polynomial-time algorithms like spectral methods or message passing.

III. CHANNEL UNIVERSALITY

Let us now show that in order to characterize the mutual information per variable, it suffices to consider an equivalent Gaussian channel. We give a detailed rigorous proof in Appendix A and present here only its main idea. We assume that the prior $p(x)$ has a finite support and denote

\[ S_{ij} = \frac{\partial}{\partial w} \log P_{\text{out}}(Y_{ij}|w)|_{w=0}, \]  
\[ S'_{ij} = \frac{\partial^2}{\partial w^2} \log P_{\text{out}}(Y_{ij}|w)|_{w=0}. \]

We assume $E_{Y_{ij}|0}[S_{ij}]$, $E_{Y_{ij}}[S'_{ij}]$, and $\partial^2_w \log P_{\text{out}}(Y_{ij}|w)|_{w=0}$ are all bounded.

Note that $W_{ij} = O(1/\sqrt{n})$. Using Taylor’s expansion of $\log P_{\text{out}}(y|w)$ at $w = 0$, for all $i \leq j$, we can write

\[ P_{\text{out}}(Y_{ij}|W_{ij}) = P_{\text{out}}(Y_{ij}|0)e^{W_{ij}S_{ij} + \frac{1}{2}W_{ij}^2S'_{ij} + O(\alpha^{-3/2})}. \]

Thus,

\[ P_{\text{out}}(Y|W) = P_{\text{out}}(Y|0)e^{\sum_{i\leq j}(W_{ij}S_{ij} + \frac{1}{2}W_{ij}^2S'_{ij}) + O(\sqrt{n})}, \]

Classical properties of the Fisher information give that $E_{Y_{ij}|0}[S_{ij}'] = -E_{Y_{ij}|0}[S_{ij}^2] = -1/\Delta$. Using the fact that $P_{\text{out}}(y|w)$ is close to $P_{\text{out}}(y|0)$, one can further argue that $E_{Y_{ij}|W_{ij}}[S'_{ij}] = -1/\Delta + O(n^{-1/2})$. By concentration inequalities, we expect that

\[ \sum_{i\leq j} W_{ij}^2 S'_{ij} \approx \sum_{i\leq j} W_{ij}^2 E_{Y_{ij}|W_{ij}}[S'_{ij}] = \sum_{i\leq j} W_{ij}^2 / \Delta + O(\sqrt{n}). \]

Thus

\[ P_{\text{out}}(Y|W) \approx P_{\text{out}}(Y|0)e^{\sum_{i\leq j}(W_{ij}S_{ij} - \frac{1}{2}W_{ij}^2) + O(\sqrt{n})}, \]

and consequently

\[ P(W|Y) \propto P(W)e^{-\frac{1}{\Delta} \sum_{i\leq j}(W_{ij}S_{ij} - W_{ij}^2) + O(\sqrt{n})}. \]

Hence, we expect that

\[ I(W;Y) = I(W;W + \sqrt{\Delta} \xi) + O(\sqrt{n}). \]

In other words, the mutual information per variable $I(x;Y)/n$ is asymptotically equal to the mutual information per variable of a Gaussian channel with noise variance $\Delta$.

IV. PROVING THE UPPER BOUND

A. Mutual information and free energy

Using the channel universality, we only need to deal with the Gaussian output. The mutual information between the observation $Y$ and the unknown vector $x$ is defined using the entropy as $I(x;Y) = H(Y) - H(Y|x)$. For the Gaussian noise, a straightforward computation shows that the mutual information per variable is expressed as

\[ \frac{I(x;Y)}{n} = f + \frac{\mathbb{E}_x(x^2)}{4\Delta}, \]

where $f = -\mathbb{E}_Y[\log Z(Y)]/n$ is the average free energy per variable using the standard statistical physics terms, and $Z(Y)$ is the partition function defined by

\[ Z(Y) \equiv \int dx p(x) \exp \left[ \sum_{i \leq j} \left( \frac{x_i^2}{2m\Delta} + \frac{x_i x_j Y_{ij}}{\sqrt{m\Delta}} \right) \right]. \]

We now show how to upper bound the free energy $f$.

B. Denoising

We first solve a simpler denoising problem. Assume we observe a noisy version of a vector $x^*$ that we denote $y$: $y = x^* + \sigma z$, where $x_i^{\text{i.i.d.}} \sim p(x)$ and $z_i^{\text{i.i.d.}} \sim \mathcal{N}(0, 1)$. The corresponding posterior distribution reads

\[ P(x|y) = \frac{1}{Z_0} p(x) \exp \left( -\frac{||x||^2}{2\sigma^2} + \frac{x \cdot y}{\sigma^2} \right). \]

For future convenience we denote the variance $\sigma^2 \equiv \frac{\Delta}{m}$, where $m$ and $\Delta$ are some so-far unspecified parameters. For this denoising problem, the averaged free energy per variables reads

\[ -nf_0 = \mathbb{E}[\log Z_0] = \mathbb{E}_{x^*,z} \left[ J \left( \frac{m}{\Delta}, \frac{m x^*}{\Delta} + \sqrt{\frac{m}{\Delta}} z \right) \right], \]

where $x^* \sim p(x)$, $z \sim \mathcal{N}(0,1)$, and $J(A, B)$ is the function defined in eq. (5). Notice how this yields a formula very close to the one in Theorem I.3.
C. The interpolation method

We now use the Guerra interpolation method, setting an artificial parameter $t$, where we interpolate between the denoising problem at $t = 0$ and the desired matrix factorization one at $t = 1$. To do so, assume that we have access to two types of noisy observations: (1) A noisy version of $x^*$, as in eq. (12), with now $\sigma^2 = \frac{\Delta}{m(1-t)}$; and (2) a noisy version of $W_{ij} = x^*_i x^*_j / \sqrt{n}$ with a Gaussian noise of variance $\Delta / t$. The posterior distribution, in this case, is given by

$$P_i(x|Y, y) = \frac{1}{Z_t} p(x) e^{\sum_{i} \left[ -\frac{x^*_i x^*_j}{2\Delta n} + (1-t) \left[ -\frac{m\|x\|^2}{2\Delta} + \frac{m\langle x, y \rangle}{\Delta} \right] \right]}.$$  

This model interpolates between the denoising problem at $t = 0$ and the one of the matrix factorization problem at $t = 1$. Using the fundamental theorem of algebra, we write

$$-n f_1 = \mathbb{E}_{x^*, z, \xi} [\log Z_t] = f_0 - \int_0^1 dt \frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t].$$  

(The free energy at $t=0$ is precisely given by (14). Using now eq. (10) and eq. (4) we write

$$I(x; Y) = ib(m) - \frac{m^2}{4\Delta} - \frac{1}{n} \int_0^1 dt \frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t].$$  

Theorem I.2 follows from the following lemma:

Lemma IV.1. For all positive $n$ and $t \in [0, 1]$, we have

$$\frac{1}{n} \frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t] \geq \frac{m^2}{4\Delta}. \quad (17)$$

D. The proof

Define

$$H_t(z, x^*, \xi, z) = \sum_{i,j} \frac{tx^*_i x^*_j}{\Delta n} + \frac{tx^*_i x^*_j x^*_i x^*_j}{\Delta n} + x^*_i x^*_j \sqrt{\frac{t}{n\Delta i,j}} \left[ (1-t) m \|x\|^2 \right] + \frac{t}{n\Delta i,j} \xi_{ij} \right] - \frac{(1-t)m\|x\|^2}{2\Delta} + \frac{(1-t)m\langle x, x^* \rangle}{\Delta} + \langle x, z \rangle \sqrt{\frac{m(1-t)}{\Delta}}.$$  

Then $P_t(x|x^*, z, \xi) = p(x) e^{H_t} / Z_t$. Now we need to compute $\frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t]$. Notice that

$$\partial_t H_t = \sum_{i,j} A_{ij} + \sum_i B_i,$$

where

$$A_{ij} = \frac{-t x^*_i x^*_j}{2\Delta n} + \frac{x^*_i x^*_j}{\Delta n} + \frac{x^*_i x^*_j}{\sqrt{\Delta n t}} \xi_{ij}, \quad (18)$$

$$B_i = \frac{m x^*_i}{2\Delta} - \frac{x^*_i}{\Delta} \frac{z_i}{\sqrt{\Delta (1-t)}}. \quad (19)$$

Since the prior $p(x)$ has a finite support $|A_{ij}|$ and $|B_i|$ are dominated by functions integrable with respect to $P_t$. Thus by the dominated convergence theorem,

$$\frac{d}{dt} \log Z_t = \mathbb{E}[\partial_t H_t] = \sum_{i,j} \mathbb{E}[A_{ij}] + \sum_i \mathbb{E}[B_i].$$

Moreover, $\mathbb{E}_t[A_{ij}]$ and $\mathbb{E}_t[B_i]$ are dominated by functions integrable with respect to the distribution of $\{x^*, \xi, z\}$. Thus again by the dominated convergence theorem,

$$\frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t] = \mathbb{E}_{x^*, z, \xi} \left[ \frac{d}{dt} \log Z_t \right]$$

$$= \mathbb{E}_{x^*, z, \xi} \left[ \sum_{i,j} \mathbb{E}_t[A_{ij}] + \sum_i \mathbb{E}_t[B_i] \right]. \quad (20)$$

We then compute $\mathbb{E}_{\xi_{ij}} \left[ \mathbb{E}_t[x_{ij}^2] \right]$ and $\mathbb{E}_{z_i} \left[ \mathbb{E}_t[x_{ij}^2] \right]$. We use the integration by parts to get rid of the $z_i$ and $\xi_{ij}$. In particular, for a standard Gaussian random variable $z$ and a continuous differentiable function $f(a)$ such that $f(a) e^{-a^2/2} \to 0$ as $a \to \infty$, we have that $E[z f(z)] = E[f'(z)]$. Notice that $P_t$ is a function of $\xi$ and $z$. Also, $\partial_t H_t = (1/t) x_i x_j$. Then $|\partial_t H_t|$ is dominated by a function integrable under $P_t$. By the dominated convergence theorem,

$$\partial_t H_t = \sqrt{\frac{t}{n\Delta}} \mathbb{E}_t[x_{ij}].$$

It follows that

$$\partial_t H_t = \sqrt{\frac{t}{n\Delta}} \left( \mathbb{E}_t[x_{ij}^2] - \mathbb{E}_t[x_{ij}]^2 \right).$$

Thus $\partial_t H_t$ is continuous in $\xi_{ij}$. Applying the integration by parts, it yields that

$$\mathbb{E}_{\xi_{ij}} \left[ \mathbb{E}_t[x_{ij}^2] \right] = \mathbb{E}_{\xi_{ij}} \left[ \partial_t \mathbb{E}_t[x_{ij}] \right]$$

$$= \sqrt{\frac{t}{n\Delta}} \mathbb{E}_{\xi_{ij}} \left[ \mathbb{E}_t[x_{ij}^2] - \mathbb{E}_t[x_{ij}]^2 \right].$$

Similarly, one can show that

$$\mathbb{E}_{z_i} \left[ \mathbb{E}_t[x_{ij}^2] \right] = \mathbb{E}_{z_i} \left[ \partial_z \mathbb{E}_t[x_{ij}] \right]$$

$$= \mathbb{E}_{z_i} \left[ \mathbb{E}_t[x_{ij}^2] - \mathbb{E}_t[x_{ij}]^2 \right].$$

It follows from (18) and (19) that

$$\mathbb{E}_{t, z}[A_{ij}] = \mathbb{E}_{t, z} \left[ \frac{x^*_i x^*_j}{n \Delta} - \frac{x^*_i x^*_j}{n \Delta} \right], \quad (21)$$

$$\mathbb{E}_{t, z}[B_i] = \mathbb{E}_{t, z} \left[ -\frac{m}{2 \Delta} x^*_i + \frac{m}{2 \Delta} x^*_i \mathbb{E}_t[x_{ij}] \right]. \quad (22)$$

Using the Nishimori identities given by Lemma A.4, we have that

$$\mathbb{E}_{x^*, z, \xi} \left[ \mathbb{E}_t[x_{ij}^2] \right] = \mathbb{E}_{x^*, z, \xi} \left[ \mathbb{E}_t[x_{ij} x_{ij}^*] \right], \quad (23)$$

$$\mathbb{E}_{x^*, z, \xi} \left[ (\mathbb{E}_t[x_{ij}^2])^2 \right] = \mathbb{E}_{x^*, z, \xi} \left[ \mathbb{E}_t[x_{ij} x_{ij}^*] \right]. \quad (24)$$

Combining (20), (21)–(22), and (23)–(24) yields that

$$\frac{1}{n} \frac{d}{dt} \mathbb{E}_{x^*, z, \xi} [\log Z_t]$$

$$= \frac{1}{2\Delta n^2} \sum_{i,j} \mathbb{E} [x_i x_j x_i x_j^*] - \frac{m}{2\Delta n} \sum_i \mathbb{E} [x_i x_i^*]$$

$$\geq \frac{m^2}{4\Delta} - \frac{m^2}{2\Delta} \geq \frac{m^2}{4\Delta},$$

where $m_t = \langle x, x^* \rangle$ with $x$ drawn from $P_t$. Lemma IV.1 readily follows.
V. PROVING THE LOWER BOUND

The proof for the lower bound also relies on the interpolation method. Again, the proof idea is inspired by [2].

A. An ad-hoc model

We shall first verify the free energy of a totally artificial model, that does not correspond to any Bayesian inference problem. Later, we will interpolate the desired free energy of eq. (4). For a fixed set of \((z, x^*)\), let

\[
\tilde{Z}_0 = \int dx p(x) e^{\frac{1}{2} \sum_{i \leq j} [x_i x_j + \sqrt{2} \tilde{z}] - \frac{\mu}{2} ||x||^2 + \sqrt{2} \tilde{z} (x, Z)}
\]

\[
= \int dx p(x) e^{\frac{1}{2} tr A (z, x^*)^2 - \frac{\mu}{2} ||x||^2 + \sqrt{2} \tilde{z} (x, Z) + O(1)},
\]

where \(O(1) = \frac{1}{2mn} ||x||^2\).

Using the Gaussian identity \(\mu e^{\frac{1}{2} ||x||^2} = \sqrt{a/\pi} \int e^{-am^2 + bm \cdot dm} \), with \(a = n/2A\) and \(b = (x, x^*)/\Delta\), we reach

\[
\tilde{Z}_0 \leq \sqrt{\frac{n}{2n \Delta}} \int dx p(x) \int dm \exp \left[ \frac{-nm^2}{2A} + \sum_{i} \mathcal{J} \left( \tilde{m}, m x_i^* + \sqrt{\tilde{m}} z_i \right) \right].
\]

Then, a naive application of the Laplace method suggests that

\[
\tilde{f}_0 := -\lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{z, x^*} \left[ \log \tilde{Z}_0 \right] \geq \min_{x^*} \left[ \frac{m^2}{2A} - \mathbb{E}_{x, z} \left\{ \mathcal{J} \left( \tilde{m}, m x^* + \sqrt{\tilde{m}} z \right) \right\} \right].
\]

A rigorous proof that \(\tilde{f}_0\) is indeed lower bounded by the above expression (which is only what we require) is presented in Appendix C, under the assumption that \(\mathbb{E}_{z, x^*} (m) = 0\) has a finite number of solutions.

B. Interpolation reloaded

The proof of the lower bound then proceed again via the interpolation method, where we interpolate between the ad-hoc model and the matrix factorization one by considering the following partition function, at fixed value of \(\{z, \xi, x^*\}\):

\[
\tilde{Z}_t = \int dx p(x) e^{\frac{1}{2} \sum_{i \leq j} \left[ -t x_i x_j + x_i x_j x_0 + \sqrt{2 \Delta} x_i \xi_j \right]}
\]

\[
\times e^{-\left(1-t\right) \left[ -\frac{\mu}{2} ||x||^2 + \sqrt{2 \tilde{z} (x, Z)} \right]}.
\]

This is again detailed in Appendix D.

ACKNOWLEDGMENTS

We thank J. Barbier, T. Lesieur, N. Macris and C. Moore for helpful discussions. Part of the research has received funding from the European Research Council under the European Union’s 7th Framework Programme (FP/2007-2013/ERC Grant Agreement 307087-SPARCES).
A. Proof of channel universality

Here we present the detailed proof of channel universality. In this proof, with a slight abuse of notation, we let \(\langle A, B \rangle := \sum_{i \leq j} A_{ij} B_{ij}\) for two symmetric matrices \(A\) and \(B\). By definition

\[
I(W; Y) = \mathbb{E}_{W,Y} \left[ \log \frac{P_{\text{out}}(Y|W)}{P(W) P_{\text{out}}(Y|W') dW'} \right].
\]

Define \(D_{ij} = S'_{ij} + 1/\Delta\) and \(H(W, Y) = \langle W, S \rangle - \frac{||W||^2}{2\Delta} + \frac{1}{2} \langle D, W \circ W \rangle\), where \(\circ\) denotes the element-wise matrix product. In view of (9),

\[
I(W; Y) = \mathbb{E}_{W,Y} \left[ \log \frac{e^{H(W,Y)}}{P(W') e^{H(W',Y)} dW'} \right] + O(\sqrt{n}).
\]

In the following, we compute \(\mathbb{E}_{W,Y} [H(W,Y)]\) and \(\mathbb{E}_{Y} \left[ \log \left( \int P(W') e^{H(W',Y)} dW' \right) \right]\) up to additive errors on the order of \(O(\sqrt{n})\).

**Lemma A.1.**

\[
\mathbb{E}_{W,Y} [H(W,Y)] = \frac{n (\mathbb{E}[x^2])^2}{4\Delta} + O(\sqrt{n}).
\]

**Proof:** Notice that

\[
\mathbb{E}_{Y|W} [\langle W, S \rangle] = \frac{n(n+1)}{2} W_{12} \mathbb{E}_{Y_{12}[W_{12}[S_{12}]
\]

\[
= \frac{n(n+1)}{2} W_{12} \mathbb{E}_{Y_{12}[S_{12}(1 + S_{12} W_{12} + O(n^{-1})]
\]

\[
= \frac{n(n+1)}{2} W_{12}^2 \mathbb{E}_{Y_{12}[S_{12}^2 + O(\sqrt{n})
\]

\[
= \frac{n+1}{2\Delta} x_1 x_2 + O(\sqrt{n}),
\]

where we used the fact that \(S'_{12} = 0\) and \(S'_{12} = 1/n\). It follows that

\[
\mathbb{E}_{Y|W} [H(W,Y)] = \frac{n+1}{2\Delta} (\mathbb{E}[x^2])^2 + O(\sqrt{n}).
\]

Also,

\[
\mathbb{E}_{Y} [\langle D, W \circ W \rangle]
\]

\[
= \frac{n(n+1)}{2} W_{12}^2 \mathbb{E}_{Y_{12}[D_{12}]
\]

\[
= \frac{n(n+1)}{2} W_{12}^2 \mathbb{E}_{Y_{12}[D_{12}(1 + O(|S_{12}| n^{-1/2})]
\]

\[
= O(\sqrt{n}),
\]

where we used the fact that \(S_{12} = 0\), and \(S_{12}^2\) and \(S_{12}^3\) are bounded. The lemma readily follows by observing that

\[
\mathbb{E} [W||2] = \sum_{i \leq j} \mathbb{E}[x_i^2 x_j^2] = \frac{n+1}{2\Delta} (\mathbb{E}[x^2])^2
\]

Define

\[
\Phi(S) = \log \left( \int P(W') e^{H(W',Y)} dW' \right).
\]

**Lemma A.2.**

\[
\mathbb{E}_{Y} \left[ \log \left( \int P(W') e^{H(W',Y')} dW' \right) \right] = \mathbb{E}_{Y} [\Phi(S)] + O(\sqrt{n}).
\]

**Proof:** Notice that

\[
|\langle D, W \circ W \rangle| = \frac{1}{n} \sum_{i \leq j} D_{ij} x_i^2 x_j^2 | \leq \frac{1}{n} \|D\| \|x\|^2 = O(\|D\|).
\]

Thus,

\[
\mathbb{E}_{Y} \left[ \log \left( \int P(W') e^{H(W',Y')} dW' \right) \right] = \mathbb{E}_{Y} [\Phi(S)] + O(\mathbb{E}_{Y}[\|D\|]).
\]

Recall that \(D_{ij} = \frac{\partial^2 \log P_{\text{out}}(Y_{ij}|0)}{\partial y_{ij}^2} + 1/\Delta\). Then \(D\) is a symmetric matrix where \(\{D_{ij}\}_{i \leq j}\) are independent and identically distributed. Moreover,

\[
\mathbb{E}_{Y} [D_{ij}] = \mathbb{E}_{Y_{ij}|0} [D_{ij}(1 + O(|S_{ij}| \sqrt{n})) = O(1/\sqrt{n}),
\]

and since \(S_{ij}'\) is bounded, \(\mathbb{E}_{Y} [D_{ij}^2] = O(1).\) By Lalata's theorem [23], \(\mathbb{E}_{Y} [D - \mathbb{E}_{Y} [D]] = O(\sqrt{n}).\) Since \(\mathbb{E}_{Y} [D]\) is \(O(\sqrt{n})\). By triangle's inequality, \(\mathbb{E}_{Y} [\|D\|] = O(\sqrt{n})\), and the lemma follows.

Finally, we apply the generalized Lindeberg principle [24, 25] to show the following lemma. Let

\[
U = \frac{W}{\Delta} + \frac{\xi}{\sqrt{\Delta}},
\]

where \(\xi_i \sim N(0, 1)\) for \(i \leq j\) and \(\xi_{ij} = \xi_{ij}\).

**Lemma A.3.**

\[
\mathbb{E}_{Y} [\Phi(S)] = \mathbb{E}_{U} [\Phi(U)] + O(\sqrt{n}).
\]

Define \(a_{ij} = \mathbb{E}[S_{ij}] - \mathbb{E}[U_{ij}]\) and \(b_{ij} = \mathbb{E}[S_{ij}^2] - \mathbb{E}[U_{ij}^2]\). Note that

\[
\mathbb{E}_{Y} [S_{ij}] = \mathbb{E}_{Y_{ij}|0} [S_{ij}(1 + W_{ij} S_{ij} + O(n^{-1}))]
\]

\[
= \frac{W_{ij}}{\Delta} + O(n^{-1}),
\]

and

\[
\mathbb{E}_{Y} [S_{ij}^2] = \mathbb{E}_{Y_{ij}|0} [S_{ij}^2(1 + O(|S_{ij}| n^{-1/2}))]
\]

\[
= \frac{1}{\Delta} + O(n^{-1/2}),
\]

where we used the fact that \(\mathbb{E}_{Y_{ij}|0} [S_{ij}^3] = O(1).\) Thus \(a_{ij} = O(n^{-1})\) and \(b_{ij} = O(n^{-1/2}).\) Also, one can check that \(\mathbb{E}[S_{ij}^3] + \mathbb{E}[U_{ij}^3] = O(1).\) Let \(f(W)\) denote the expectation of \(f(W)\) with respect to the measure defined
by \( P(W)e^{\mathcal{H}(W,Y)}dW / \int P(W')e^{\mathcal{H}(W',Y)}dW' \). It follows that

\[
\frac{\partial \Phi}{\partial S_{ij}} = [W_{ij}] = O(n^{-1/2})
\]

\[
\frac{\partial^2 \Phi}{\partial S_{ij}^2} = [W_{ij}^2] - [W_{ij}]^2 = O(n^{-1})
\]

\[
\frac{\partial^3 \Phi}{\partial S_{ij}^3} = [W_{ij}^3] - 3[W_{ij}][W_{ij}^2] + 2[W_{ij}]^3 = O(n^{-3/2}).
\]

Therefore, by Lindeberg principle [24, Theorem 1.1],

\[
|E_Y[\Phi(S)] - E_U[\Phi(U)]| \leq O(n^{-1/2}) \sum_{i \leq j} a_{ij} + O(n^{-1}) \sum_{i \leq j} b_{ij} + O(\sqrt{n})
\]

\[
= O(\sqrt{n}).
\]

In conclusion, we have shown that

\[
I(W; Y) = \frac{n}{4\Delta} \left( \frac{E[z^2]}{2} - E_U[\Phi(U)] \right) + O(\sqrt{n})
\]

\[
= I(W; W + \sqrt{\Delta} \xi) + O(\sqrt{n}).
\]

\[\text{B. Nishimori identities}\]

A key ingredient in our interpolation proof is the following Nishimori identities [8, 13, 15], which hold for Bayesian inference problems.

**Lemma A.4** (Nishimori identities). Let \( x^* \) denote a random sample from a prior distribution \( p(x) \), and we observe \( y \) randomly generated from \( p(y|x^*) \). Let \( x \) and \( x' \) denote two independent random samples from \( p(x|y) \) with the posterior distribution \( p(x|y) = p(x)p(y|x)/p(y) \). Then for all \( f \) such that \( E[|f(x, x^*)|] < \infty \),

\[
E[f(x, x^*)] = E[f(x, x^*)].
\]

**Proof:** By definition,

\[
E[f(x, x^*)] = \int p(x^*) \int p(y|x^*) f(x, x^*) p(x|y) dx dy dx^*
\]

\[
= \int p(y) \int f(x, x^*) p(x|y) p(x^*|y) dx dx^* dy
\]

\[
= E[f(x, x')],
\]

where \((a)\) follows from the fact that \( p(x^*)p(y|x^*) = p(x^*)p(y) \) and Fubini’s theorem. \( \Box \)

\[\text{C. Laplace method}\]

We present the rigorous proof. For fixed \( x^* \) and \( z \), let

\[
G(m, x^*, z) = \frac{-m^2}{2\Delta} + \frac{1}{n} \sum_i \mathcal{J} \left( \frac{m}{\Delta}, \frac{mx^*}{\Delta} + \sqrt{\frac{m}{\Delta}} z_i \right).
\]

and

\[
\eta(m, x^*_i, z_i) = \Delta \times \partial_m \mathcal{J} \left( \frac{m}{\Delta}, \frac{mx^*_i}{\Delta} + \sqrt{\frac{m}{\Delta}} z_i \right)
\]

\[
= [x_i]x^*_i;
\]

where \([x_i]\) denotes the mean of \( x_i \) under the distribution proportional to \( \exp \left( -\frac{m^2}{2\Delta} + x_i \frac{m}{\Delta} + \sqrt{\frac{m}{\Delta}} z_i \right) p(x) dx \).

Since \( p(x) \) has a finite support, \( \eta \) is bounded. Without loss of generality, assume \( \|\eta\| \leq C \). It follows that \( m \) achieving the maximum value of \( G(m, x^*, z) \) must satisfy

\[
m = \frac{1}{n} \sum_i \eta(m, x_i, z_i).
\]

Similarly, \( m \) achieving the maximum value of \( E[G(m, x^*_1, z_1)] \) must satisfy

\[
m = E[\eta(m, x^*_1, z_1)].
\]

Let \( S \) denote the set of the solutions in \([-C, C]\) of the above fixed point equation. By assumption, \( \eta^* \in S \) is non-empty for all \( \delta \). Thus, \( \sup_{m \in S} \eta(m, x_i, z_i) \) has a finite support, \( \eta \) is bounded. Without loss of generality, assume \( \|\eta\| \leq C \). It follows that \( m \) achieving the maximum value of \( \max_{m \in S} G(m, x^*, z) \) must satisfy

\[
m = \frac{1}{n} \sum_i \eta(m, x_i, z_i).
\]

Notice that

\[
\max_{m \in S} G(m, x^*, z) \leq \max_{m \in S} \left| G(m, x^*, z) - E[G(m, x^*, z)] \right| + \max_{m \in S} E[G(m, x^*, z)].
\]
Recall that for any fixed $m \in S$,
\[
\left| G(m, x^*, z) - \mathbb{E}[G(m, x^*, z)] \right| = \frac{1}{n} \sum_i \mathcal{J} \left( \frac{\hat{m}}{\Delta} \frac{mx_i^*}{\Delta} + \sqrt{\frac{\hat{m}}{\Delta} z_i} \right)
- \mathbb{E} \left[ \mathcal{J} \left( \frac{\hat{m}}{\Delta} \frac{mx_i^*}{\Delta} + \sqrt{\frac{\hat{m}}{\Delta} z_i} \right) \right] := T(m).
\]

Using the fact that $|x_i^*|$ is bounded, one can check that $\mathcal{J} \left( \frac{\hat{m}}{\Delta} \frac{mx_i^*}{\Delta} + \sqrt{\frac{\hat{m}}{\Delta} z_i} \right)$ is sub-Gaussian with $O(1)$ sub-Gaussian norm. Thus by Chernoff’s bound, for any fixed $m \in S$ and $t \geq 0$,
\[
\mathbb{P} \{ T(m) > t \} \leq e^{-\Omega(nt^2)}.
\]

By a union bound, it follows that with probability at most $|S|e^{-\Omega(nt^2)}$, $\max_{m \in S} T(m) > t$. It follows that
\[
\mathbb{E} \left[ \max_{m \in S} T(m) \right] = \int_0^\infty \mathbb{P} \left\{ \max_{m \in S} T(m) > t \right\} dt
\leq \int_0^\delta dt + \int_\delta^\infty \mathbb{P} \left\{ \max_{m \in S} T(m) > t \right\} dt
\leq \delta + |S| \int_\delta^\infty e^{-\Omega(nt^2)} = o(1).
\]

In view of (29), we have that
\[
\mathbb{E} \left[ \max_{m \in S} G(m, x^*, z) \right] = \max_{m \in S} \mathbb{E} \left[ G(m, x^*, z) \right] + o(1).
\]

Combining the above display with (28), it yields that
\[
\lim_{n \to \infty} \frac{1}{n} \mathbb{E} \left[ \log Z_0 \right] = \max_{m \in S} \mathbb{E} \left[ G(m, x^*, z) \right] \leq \max_{m \in S} \mathbb{E} \left[ G(m, x^*, z) \right]
= \max_m \left\{ \frac{-m^2}{2\Delta} + \mathbb{E}_{x,z} \mathcal{J} \left( \frac{\hat{m}}{\Delta} \frac{mx^*}{\Delta} + \sqrt{\frac{\hat{m}}{\Delta} z} \right) \right\},
\]

which completes the proof.

**D. Interpolating the lower bound**

Let us repeat the interpolation strategy of sec. IV-C. We start with eq. (27). Computing the derivative with respect to $t$, we find that
\[
\frac{d}{dt} \mathbb{E}_{x^*, z, \xi} \left[ \log Z_t \right] = \mathbb{E}_{x^*, z, \xi} \left[ \sum_{i \leq j} E_t[A_{ij}] + \sum_i E_t[B_i] \right],
\]

where
\[
A_{ij} = -\frac{x_i^2 x_j^2}{2\Delta n} + \frac{x_i x_j}{2\sqrt{\Delta n}} \xi_{ij}
\quad \text{ and } \quad
B_i = \frac{\hat{m}}{2\Delta} x_i^2 - 1 \left\lfloor \sqrt{\frac{\hat{m}}{\Delta (1-t)}} x_i z_i \right\rfloor.
\]

Performing again the integration by part leads to
\[
\mathbb{E}_{t, \xi} [A_{ij}] = -\mathbb{E}_{t, \xi} \left[ \frac{x_i x_j E_t[x_i x_j]}{2n\Delta} \right]
\quad \text{ and } \quad
\mathbb{E}_{t, \xi} [B_i] = \mathbb{E}_{t, z} \left[ \frac{x_i E_t[x_i] \hat{m}}{2\Delta} \right].
\]

Let $x'$ be an independent copy of $x$. Therefore we have
\[
\frac{1}{n} \frac{d}{dt} \mathbb{E}_{\xi, z, x^*, x} \left[ \log Z_t \right]
= \frac{1}{2n^2 \Delta} \sum_{i \leq j} \mathbb{E}[x_i x_j x_i' x_j'] - \frac{\hat{m}}{2\Delta} \sum_i \mathbb{E}[x_i x_i']
\geq \frac{1}{4n^2 \Delta} \mathbb{E}[\langle x, x' \rangle^2] - \frac{\hat{m}}{2\Delta} \frac{\hat{m}}{4\Delta} \geq -\frac{\hat{m}^2}{4\Delta},
\]

which, with (15) and (26), leads to Theorem (I.3).